2014

Incorporating Dependence Boundaries in Simulating Associated Discrete Data

Mary E. Haynes
Virginia Commonwealth University, haynesme2@vcu.edu

INCORPORATING DEPENDENCE BOUNDARIES IN SIMULATING ASSOCIATED DISCRETE DATA

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

MARY EMILIA HAYNES

Bachelor of Arts with Bryan College - August 2005 to May 2009

Director: Roy T. Sabo, Ph.D.

Associate Professor, Department of Biostatistics

Virginia Commonwealth University

Richmond, Virginia

December, 2014

Acknowledgements

I would like to thank my husband, Caleb, for his undying devotion, love, support, and patience during the time it has taken me to complete this program. I would like to thank my parents for encouraging me to finish, and especially to my father. I would like to thank my advisor, Dr. Sabo, for his excellence in teaching and guiding me through all parts of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

INCORPORATING DEPENDENCE BOUNDARIES IN SIMULATING ASSOCIATED DISCRETE DATA

By Mary Emilia Haynes

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2014.

Director: Roy T. Sabo, Ph.D.
Associate Professor, Department of Biostatistics

In the study of associated discrete variables, limitations on the range of the possible association measures (Pearson correlation, odds ratio, etc.) arise from the form of the joint probability function between the variables. These limitations are known as the Fréchet bounds. The bounds for cases involving associated binary variables are explored in the context of simulating datasets with a desired correlation and set of marginal probabilities. A new method for creating such datasets is compared to an existing method that uses the multivariate probit. A method for simulating associated binary variables using a desired odds ratio and known marginal probabilities is also presented. The Fréchet bounds for correlation between dependent binomial and negative binomial variables are determined as families of ranges in various cases. An example of a realistic analysis involving the Fréchet bounds in a dependent binomial setting is presented.

CHAPTER 1

INTRODUCTION

Many significant contributions to mathematics, statistics, and probability were made by Maurice René Fréchet in the early and middle 20th century. Among these contributions was a discovery of bounds on the probabilities of conjunctions ("and" operations) and disjunctions ("or" operations) [7] [8]. Focusing on the probabilities of conjunctions, suppose A_i (i=1, 2, ... , n) are events. The Fréchet bounds on the conjunction of all events $P(A_1 \& A_2 \& ... \& A_n)$ are as follows:

$$\max[P(A_1)+P(A_2)+...+P(A_n)-(n-1),0] \leq P(A_1 \& A_2 \& ... \& A_n) \leq \min[P(A_1), P(A_2), ..., P(A_n)]$$

These bounds have implications for various probabilistic and statistical methods currently in use, particularly for non-normal, discrete, dependent data. Normal data have many appealing qualities and are generally easy to work with, even when dependent normal random variables are being analyzed. Non-normal data–especially discrete distributions such as the binary, binomial, negative binomial, and Poisson distributions–come with less flexibility and more restrictions on the possible outcomes. The Fréchet bounds arise from the restrictive forms of the probability mass functions (pmfs) based on the assumed marginal parameters of the discrete dependent variables in question. Most measures of association between discrete dependent variables will have limitations. The Pearson correlation and the odds ratio are two of the most common measures of association used in statistical methods.

Many common modeling and simulation techniques overlook the Fréchet bounds, specifically those that lack a fully specified likelihood. The use of the Generalized Estimating Equations (GEE) developed by Liang and Zeger in their 1986 paper [9] would be one example of such a technique. In their 2010 paper, Sabo and Chaganty [10] showed that failing to incorporate the Fréchet bounds into an analysis can create problems in the results of models involving GEE, including incorrect parameter estimates, standard errors on the estimates, and inference (p-values). In principle, these

problems can occur whenever the Fréchet bounds are not directly accounted for, though some methodologies will perform better than others. Therefore, understanding the Fréchet bounds and using methods that incorporate or do not violate them should be important to the statistical modeler.

The Fréchet bounds on the associations take on a general form. For any two marginal distributions with cumulative distribution functions (cdfs) $F(y_i)$ and $F(y_j)$ for random variables $\mathrm{Y_i}$ and $\mathrm{Y_j}$ with outcomes y_i and y_j, respectively, the following limits on the joint distribution function $F(y_i, y_j)$ apply. Following directly from the bounds on the probability of a conjunction above:

$$F_L(y_i, y_j) = \max[F(y_i) + F(y_j) - 1, 0] \leq F(y_i, y_j) \leq \min[F(y_i), F(y_j)] = F_U(y_i, y_j) \tag{1.1}$$

The Fréchet bounds $[\rho_{ijL}, \rho_{ijU}]$ on the Pearson correlation ρ_{ij} are then defined as:

$$\rho_{ijL} = \frac{E_L(y_i, y_j) - E(y_i)E(y_j)}{\sqrt{V(y_i)V(y_j)}} \tag{1.2}$$

$$\rho_{ijU} = \frac{E_U(y_i, y_j) - E(y_i)E(y_j)}{\sqrt{V(y_i)V(y_j)}} \tag{1.3}$$

where $E(y_j)$ and $V(y_j)$ are the expected value and variance for $\mathrm{Y_j}$, respectively, and

$$E_L(y_i, y_j) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \max\left[P(y_i \geq k) + P(y_j \geq l) - 1, 0\right]$$

$$E_U(y_i, y_j) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \min\left[P(y_i \geq k), P(y_j \geq l)\right]$$

from Chaganty and Mav (2007) [2]. Note that for distributions with finite support, the sum will be limited by the upper bound of the support rather than infinity. Note also that the sums may be replaced with integrals for continuous distributions.

Chapter 2 will focus on simulating dependent binary variables, comparing the Emrich and Piedmonte [6] (EP) technique to a new "multinomial sampling" (MS) technique as to how well each performs in creating dependent variables with a specified correlation within the Fréchet bounds. The techniques will be compared based on mean observed correlation, observed standard deviation of the mean correlation, the proportion of simulation runs which had correlations falling within the Fréchet bounds, and the difference between the mean observed correlation and the spec-

ified correlation (bias). Using the same EP and MS techniques, a two-group pre-/post-treatment comparison study will be simulated using both two and three repeated measures to determine which simulation technique can better replicate the correlations and inference desired. The empirical correlation and marginal probabilities will be calculated and examined. GEE models will be used to estimate working correlation and inference.

Chapter 3 will cover calculation of the Fréchet bounds in the three-variable dependent binary case using unstructured and common odds ratio association structures. The Fréchet bounds [ψ_{ijL}, ψ_{ijU}] on the odds ratio ψ_{ij} are not as restrictive as those on the correlation in that the bounds do not affect the range of the odds ratio unless there are at least three associated random variables Y_i, Y_j, and Y_k. In a three-variable case, there are three odds ratios that could be calculated, and only one of those odds ratios will be restricted by the Fréchet bounds. Certain features of the multinomial sampling method in simulating dependent binary variables with specified marginal probabilities and odds ratios will be examined, including: the proportion of simulations where all odds ratios fall within the Fréchet bounds, the proportion of simulations needing an adjustment to the odds ratio (to account for empty cells), the difference between the estimated odds ratio and the specified odds ratio (bias), and the standard deviation of the odds ratio.

Chapter 4 will explore the Fréchet bounds in select binomial and negative binomial distributions, examining two- and three-variable cases to observe how the Fréchet bounds change given different combinations of the marginal probabilities and other parameters. Figures displaying the sets of bounds will be used extensively, and descriptions of interesting features of the figures will accompany them. As a realistic example, the Fréchet bounds will be calculated and compared to correlation results from an analysis using data from the National Health and Nutrition Examination Survey (NHANES). The empirical correlation and marginal probabilities will be calculated and used to calculate the appropriate Fréchet bounds. A GEE model will be used to estimate the working correlation to see whether it falls within the bounds.

The final chapter, Chapter 5 will summarize the findings of the previous chapters, discuss the limitations of this work, and describe immediate extensions of this project.

CHAPTER 2

SIMULATING DEPENDENT BINARY VARIABLES USING THE CORRELATION

2.1 Introduction

The simulation of dependent binary random variables is a necessary component of many types of research. One of the most common methods currently used was developed by Emrich and Piedmonte [6] and is based on a bivariate standard normal variable to simulate the binary data. A new method called "multinomial sampling" will be introduced and compared to the Emrich and Piedmonte method using a variety of measures including the mean observed correlation and its observed standard deviation, the proportion of simulation runs which have correlations falling within the Fréchet bounds, and the difference between the mean observed correlation and the specified correlation (bias). The methods will be compared in two- and three-variable dependent binary cases with various values for the marginal probabilities of the variables.

The Emrich and Piedmonte and multinomial sampling methods will also be compared by simulating binary data for a pre-treatment test and both one and two post-treatment tests for two cases of a general two-group comparison study. The groups will be simulated to have different changes in success rates (i.e. marginal probabilities) from the pre-treatment test to the post-treatment test(s). The Generalized Estimating Equations as developed by Liang and Zeger [9] will be used for testing whether the change in success rate differs between the two groups. The p-value will be recorded as well as the working correlation and whether the working correlation falls within the Fréchet bounds, and these will be compared between the simulation methods.

2.2 The Fréchet Bounds on the Correlation

Let Y_1, Y_2, ..., Y_J be random variables from discrete distributions which the correlated outcomes y_1, y_2, ..., y_J are to be generated, and let the correlation between any two distinct Y_i and Y_j

(i, j = 1, 2, ..., J; i $\neq$ j) be denoted as ρ_{ij}.

Recall the definition of the Fréchet bounds in Chapter 1, particularly Equations 1.1, 1.2, and 1.3. Chaganty and Joe [1] describe these bounds for the two- and three-variable binary cases. Let p_j be the marginal probability of its respective Y_j, and $q_j = 1 - p_j$. For the two-variable case (i.e. J = 2),

$$\rho_{12L} = \max\left[-\left(\frac{p_1 p_2}{q_1 q_2}\right)^{1/2}, -\left(\frac{q_1 q_2}{p_1 p_2}\right)^{1/2}\right]$$
$$\rho_{12U} = \min\left[\left(\frac{p_2 q_1}{p_1 q_2}\right)^{1/2}, \left(\frac{p_1 q_2}{p_2 q_1}\right)^{1/2}\right]$$

For J $>$ 2, the Fréchet bounds change according to the chosen correlation structure. The compound symmetric (CS) and first-order auto-regressive structures (AR(1)) will be examined for J=3. The Fréchet bounds for these situations are as follows.

In the CS case for J=3, $\rho_{12} = \rho_{13} = \rho_{23} = \rho$, so the bounds are denoted as $\rho_{L,CS}$ and $\rho_{U,CS}$ for the lower and upper bounds, respectively. Then,

$$\rho_{L,CS} = \max\left[-\frac{(p_1 p_2 p_3 + q_1 q_2 q_3)}{\sigma_1\sigma_2 + \sigma_1\sigma_3 + \sigma_2\sigma_3}, \max\left[-\left(\frac{p_i p_j}{q_i q_j}\right)^{1/2}, -\left(\frac{q_i q_j}{p_i p_j}\right)^{1/2}; 1 \leq i < j \leq 3\right]\right]$$
$$\rho_{U,CS} = \min\left[\min\left[-\left(\frac{p_i q_j}{p_j q_i}\right)^{1/2}, -\left(\frac{p_j q_i}{p_i q_j}\right)^{1/2}; 1 \leq i < j \leq 3\right]\right]$$

where σ_j is the standard deviation of Y_j.

In the AR(1) case for J=3, $\rho_{12} = \rho_{23} = \rho$ and $\rho_{13} = \rho^2$. Since the ρ parameter is common to all three correlation estimates, even though ρ_{13} estimates the square, the bounds on ρ are denoted as $\rho_{L,AR(1)}$ and $\rho_{U,AR(1)}$ for the lower and upper bounds, respectively. Then,

$$\rho_{L,AR(1)} = \max\left[\max\left[-\left(\frac{p_i p_{i+1}}{q_i q_{i+1}}\right)^{1/2}, -\left(\frac{q_i q_{i+1}}{p_i p_{i+1}}\right)^{1/2}; i = 1,2\right]\right]$$
$$\rho_{U,AR(1)} = \min\left[\min\left[-\left(\frac{p_i q_{i+1}}{q_i p_{i+1}}\right)^{1/2}, -\left(\frac{q_i p_{i+1}}{p_i q_{i+1}}\right)^{1/2}; i = 1,2\right]\right].$$

2.3 Basic Algorithms for Simulating Binary Correlated Data

2.3.1 Emrich and Piedmonte (EP) Method

Emrich and Piedmonte [6] use a multivariate probit approach to simulate dependent binary variables. Let $\Phi[x_1, x_2, r]$ represent the bivariate standard normal cumulative density function (cdf) for dependent normal random variables X_1 and X_2 that have correlation r. In order to generate the binary correlated outcomes, first solve the equations below for r_{ij}.

$$\Phi\left[z(p_i), z\left(p_j\right), r_{ij}\right] = \rho_{ij}\left(p_i q_i p_j q_j\right)^{1/2} + p_i p_j \tag{2.1}$$

where $z(p)$ denotes the pth quantile of the standard normal distribution. The quantities ρ_{ij}, p_i, p_j, q_i, and q_j are known. Once each r_{ij} has been solved for by using some numerical technique (in this case, a vector was created containing outcomes for the absolute difference between the right-hand and left-hand sides of the equation for values of r_{ij} from -0.999 to 0.999 in increments of 0.001, and the r_{ij} corresponding to the value nearest zero was chosen, see Appendix A.1), the next step is to generate a J-dimensional multivariate normal random variable, $\mathbf{Z} = (Z_1, ..., Z_J)^T$ with mean $\mathbf{0}$ and correlation matrix $\Sigma = \left(\left(r_{ij}\right)\right)$. (This is readily accomplished using the RANDNORMAL function in SAS.) Finally, for j = 1, ..., J, set $Y_j = 1$ if $Z_j \leq z\left(p_j\right)$, and set to 0 otherwise. This gives a matrix of dependent binary outcomes, and the algorithm may be iterated in order to create n observations.

2.3.2 Multinomial Sampling (MS) Method

The multinomial sampling method uses a multinomial distribution on the possible dependent binary outcomes that can be created through the joint and marginal probabilities, along with the desired correlation. Let p_{ij} represent the joint probability of Y_i and Y_j, and let p_{ijk} represent the joint probability of Y_i, Y_j, and Y_k.

Given a correlation ρ_{ij}, first solve the equations for the two-variable joint probabilities.

$$p_{ij} = p_i p_j + \rho_{ij}\sqrt{p_i q_i}\sqrt{p_j q_j} \tag{2.2}$$

The joint probability for three or more variables is not fully defined by the marginal probabilities and the correlation. Therefore, the MS method finds the minimum and maximum p_{ijk} and uses the average of the two in order to define the joint probability for three variables. That is,

$$p_{ijk,L} = \max\left\{0,\, p_{ij}+p_{ik}-p_i,\, p_{ij}+p_{jk}-p_j,\, p_{ik}+p_{jk}-p_k\right\}$$

$$p_{ijk,U} = \min\left\{p_{ij},\, p_{ik},\, p_{jk},\, 1-p_i-p_j-p_k+p_{ij}+p_{ik}+p_{jk}\right\}$$

$$p_{ijk} = \frac{p_{ijk,L}+p_{ijk,U}}{2}$$

Other methods for choosing p_{ijk} could be used, however, this method is intuitive and simple. If more than three correlated binary variables are being simulated, the joint probability of the higher-order combinations of variables will need to be calculated.

Using these quantities, the pdf and cdf of the joint distribution are calculated. The cdf is created by progressively summing the pdf as seen in Table 2.1. After the cdf is determined, simulation can begin. A Uniform(0,1) random variable, U, is simulated, and the observation is categorized by a decision rule based on the "steps" of the cdf. For example, if $P_{11} < U \leq P_{11}+P_{10}$, then the observation is recorded as $y_1 = 1$ and $y_2 = 0$ or simply as 10. This is similar in the three-variable case.

Table 2.1. Two-Variable Joint Probabilities, pdf, cdf, Decision Rules, and Outcomes for Simulation Purposes

Joint pdf	cdf	Decision Rule	Recorded Outcome
$P_{11} = p_{12}$	P_{11}	$U \leq P_{11}$	11
$P_{10} = p_1 - p_{12}$	$P_{11}+P_{10}$	$P_{11} < U \leq P_{11}+P_{10}$	10
$P_{01} = p_2 - p_{12}$	$P_{11}+P_{10}+P_{01}$	$P_{11}+P_{10} < U \leq P_{11}+P_{10}+P_{01}$	01
$P_{00} = 1-p_1-p_2+p_{12}$	$P_{11}+P_{10}+P_{01}+P_{00}$	$U > P_{11}+P_{10}+P_{01}$	00

where the subscripts on P indicate whether each binary outcome is successful, with 1 for success

and 0 for failure. For example, P_{01} is the probability that Y_1 failed and Y_2 succeeded.

Similarly, for the three-variable case, the joint pdf is as follows:

$$
\begin{aligned}
P_{111} &= p_{123} \\
P_{110} &= p_{12} - p_{123} \\
P_{101} &= p_{13} - p_{123} \\
P_{011} &= p_{23} - p_{123} \\
P_{100} &= p_1 - p_{12} - p_{13} + p_{123} \\
P_{010} &= p_2 - p_{12} - p_{23} + p_{123} \\
P_{001} &= p_3 - p_{13} - p_{23} + p_{123} \\
P_{000} &= 1 - p_1 - p_2 - p_3 + p_{12} + p_{13} + p_{23} - p_{123}
\end{aligned}
$$

where the subscripts on P are defined as above. The decision rules and outcomes would be similar to those in the two-variable case as described in Table 2.1.

The algorithm iterates until n observations are reached, giving a matrix of correlated binary variables.

2.4 Simulation Study

The performance of the EP and MS methods were compared with respect to: the proportion of simulations for which the estimated correlation is within the admissible range, the bias, and the efficiency as represented by standard deviation of the estimated correlation. The sample size was fixed at 50 subjects, and the results are presented for a wide range of correlations within the Fréchet bounds. Two-measure cases consisted of all cases in which p_1 and p_2 vary from 0.1 to 0.9 in increments of 0.2 where $p_1 \leq p_2$. Three-measure cases will consist of select cases due to the large number of possible cases. Both CS and AR(1) structures will be examined. In the three-measure cases, the estimated correlation between each of the three pairs of variables will be compared between the methods separately, although they are estimates of the same quantity.

For example, theoretically, $\hat{\rho}_{12} = \hat{\rho}_{13} = \hat{\rho}_{23} = \rho$ in the CS case, but in the actual simulations, the estimates are likely to differ. Thus, a comparison between the EP and MS methods will be made for each $\hat{\rho}_{ij}$ estimate.

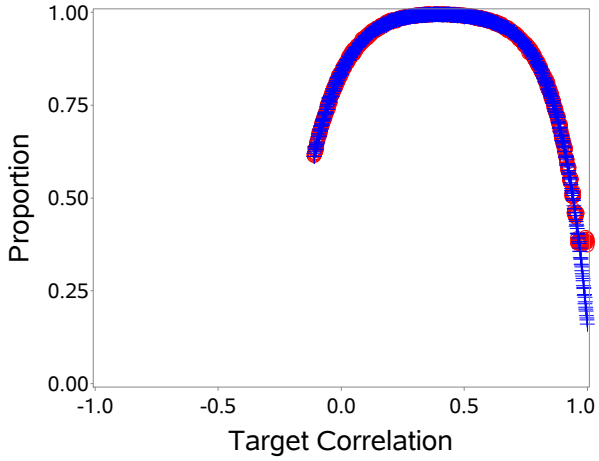
2.4.1 Methods

In all cases described above, 1000 evenly-spaced target correlations within and including the estimated Fréchet bounds were compared between the EP and MS methods. In simulating the binary data, 50 observations with 10,000 iterations were used to estimate the properties in question. The random seed chosen for the simulations was 47. The algorithms described in Section 2.3 were used to compute the simulations. The observed variance of each binary vector was calculated; any runs which had zero variance were not used in calculating the mean observed correlation or the percent of observed correlations. The observed correlations were estimated using the Pearson correlation coefficient and categorized as to whether they fell within the Fréchet bounds. The mean correlation ($\hat{\rho}$) was calculated and will be used in the presentation of results as an estimate for the consistency. The observed standard deviation was calculated as well as the proportion of runs which had correlations falling within the bounds. The bias was calculated by subtracting the target correlation from the calculated correlation (ρ). Where symmetry between cases is mentioned, it refers to rotation about the target correlation zero. All calculations were completed using SAS 9.4 (The SAS Institute, Cary, NC). PROC IML was used for the simulations, and PROC GPLOT was used for plotting results.

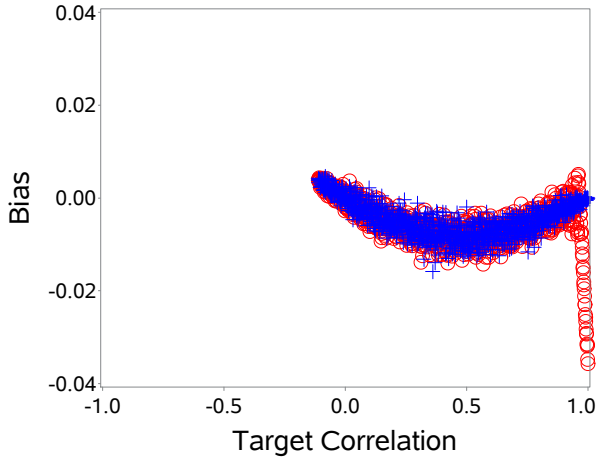
2.4.2 Results

2.4.2.1 Two-Measure Cases

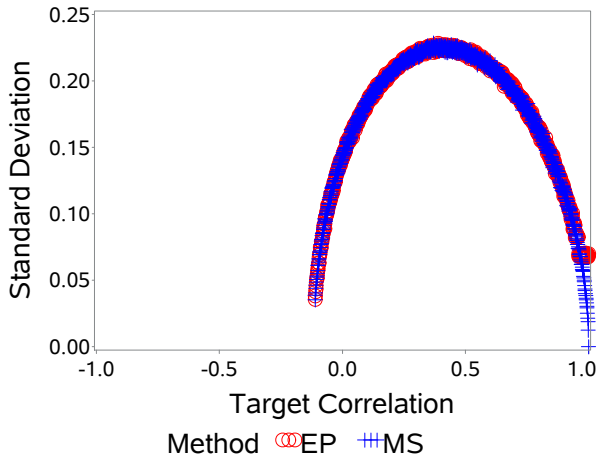
The order of p_1 and p_2 is of no consequence in the two-variable case. As such, certain cases have been eliminated due to redundancy (e.g. the case where $p_1 = 0.3$ and $p_2 = 0.1$).



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.1. Case: $p_1 = 0.1$, $p_2 = 0.1$

Case: $p_1 = 0.1$, $p_2 = 0.1$

The Fréchet range for this case is [-0.111, 1.000]. The proportion of simulations falling within the Fréchet bounds in the MS case varied between 0.612 at the lower bound to nearly 1 (>0.996) in the middle of the range then back down to 0.160. In the EP case, the proportion varied between 0.614 at the lower bound to nearly 1 (>0.997) in the middle of the range, back to 0.382 at the upper bound.

The bias for the MS method was close to zero for all points, with a somewhat parabolic shape, going from 0.004 to a minimum of -0.016 then increasing to 0. The bias for the EP method was more extreme at the upper bound of 1, at -0.036, staying near zero otherwise. Being just over twice the maximum bias of the MS method, this indicates that the EP method does not estimate the correlation near the upper Fréchet bound well.

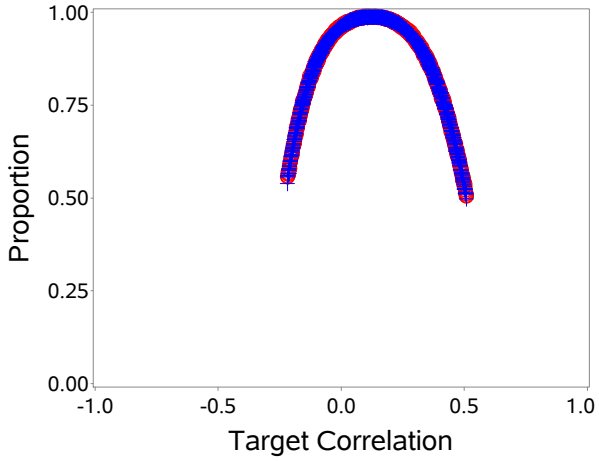
The standard deviation has a parabolic shape, as would be expected due to the nature of the estimates for the target correlation. At the lower bound, both methods were calculated to be 0.035, increasing to 0.228 in the center, then decreasing to zero in the MS case at the upper bound of 1, while only decreasing to 0.070 in the EP case, again due to the poor estimation of the target correlation at 1.

Case: $p_1 = 0.1$, $p_2 = 0.3$

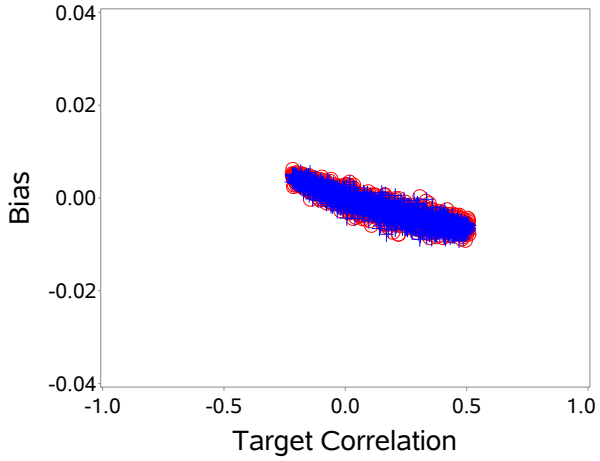
The Fréchet range for this case is [-0.218, 0.509]. The proportion of simulations falling within the Fréchet bounds in the MS case varied between 0.539 at the lower bound to nearly 1 (>0.991) in the middle of the range then back down to 0.497. In the EP case, the proportion varied between 0.557 at the lower bound to nearly 1 (>0.991) in the middle of the range, back to 0.507 at the upper bound.

The bias for both methods was comparable, staying near zero throughout the Fréchet range, with the appearance of a slight negative incline, going from near 0.004 to about -0.006 for both methods.

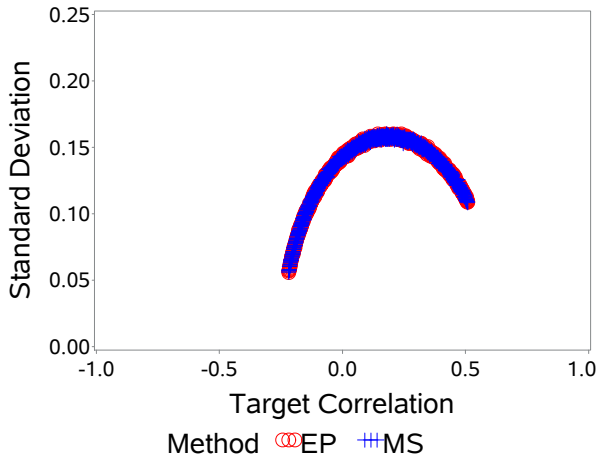
At the lower bound, the standard deviation for both methods was calculated to be 0.056,



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation

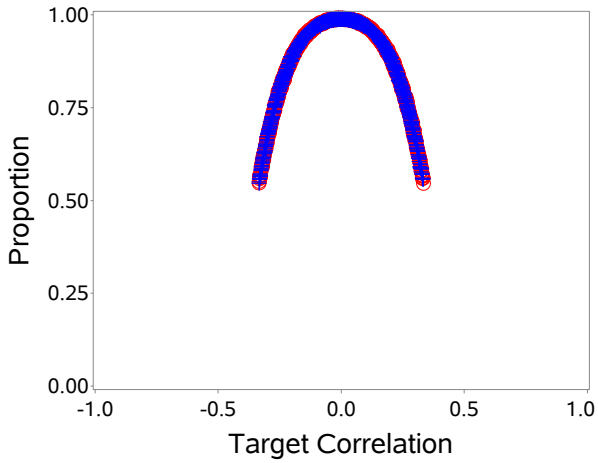


(c) Standard deviation of the estimated correlation

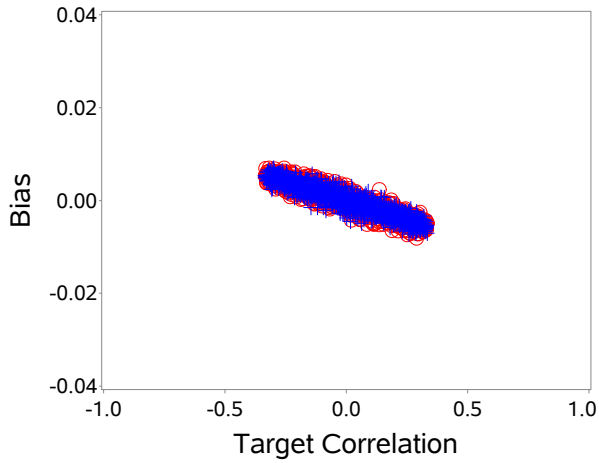
Fig. 2.2. Case: $p_1 = 0.1$, $p_2 = 0.3$

increasing to 0.160 in the center, then decreasing to 0.109 at the upper bound.

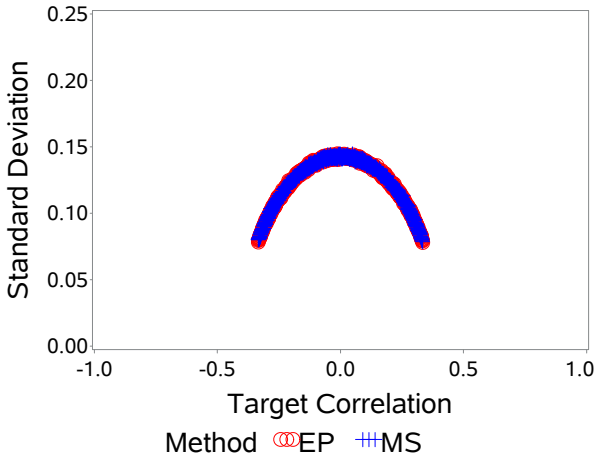
Case: $p_1 = 0.1$, $p_2 = 0.5$



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.3. Case: $p_1 = 0.1$, $p_2 = 0.5$

The Fréchet range for this case is [-0.333, 0.333]. The proportion of simulations falling within the Fréchet bounds in the MS case varied between 0.550 at the lower bound to nearly 1 (>0.992) in the middle of the range then back down to 0.559, nearly symmetrical. In the EP case, the proportion varied between 0.547 at the lower bound to nearly 1 (>0.992) in the middle of the range, back to

0.550 at the upper bound, again, nearly symmetrical.

The bias for both methods was comparable, staying near zero throughout the range with an appearance of a slight negative incline, going from approximately 0.005 for both methods to -0.007 for the MS method and 0.005 for the EP method.

At the lower bound, the standard deviation for both methods was calculated to be 0.080, increasing to 0.145 in the center, then decreasing to 0.078 at the upper bound.

Case: $p_1 = 0.1$, $p_2 = 0.7$

The Fréchet range for this case is [-0.509, 0.218]. The proportion of simulations falling within the Fréchet bounds in the MS case varied between 0.506 at the lower bound to nearly 1 ($>$0.991) in the middle of the range then back down to 0.549. In the EP case, the proportion varied between 0.512 at the lower bound to nearly 1 ($>$0.990) in the middle of the range, back to 0.559 at the upper bound. Note that this case is nearly symmetric to the case where $p_1 = 0.1$, $p_2 = 0.3$.

The bias for both methods was comparable, staying near zero throughout the range with an appearance of a slight negative incline, going from approximately 0.007 to -0.004.

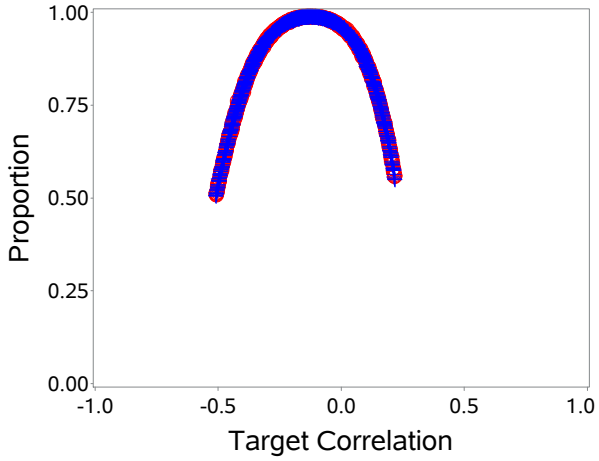
At the lower bound, the standard deviation for both methods was calculated to be 0.108, increasing to 0.160 in the center, then decreasing to 0.056 at the upper bound.

Note that this case is symmetric to the case where $p_1 = 0.1$, $p_2 = 0.3$.

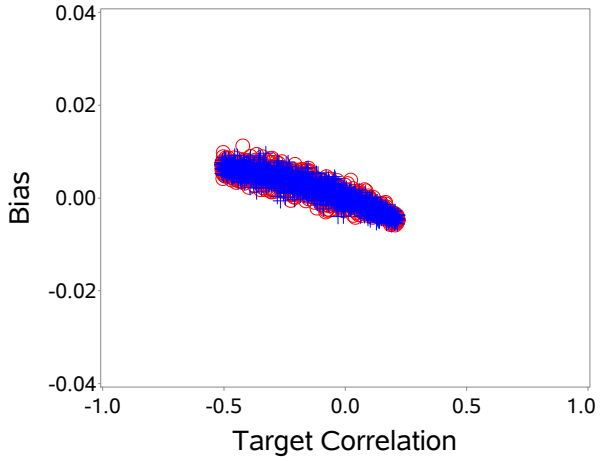
Case: $p_1 = 0.1$, $p_2 = 0.9$

The Fréchet range for this case is [-1.000, 0.111]. The proportion of simulations falling within the Fréchet bounds in the MS case varied between 0 at the lower bound to nearly 1 ($>$0.996) in the middle of the range then back down to 0.611. In the EP case, the proportion varied between 0.267 at the lower bound to nearly 1 ($>$0.996) in the middle of the range, back to 0.613 at the upper bound.

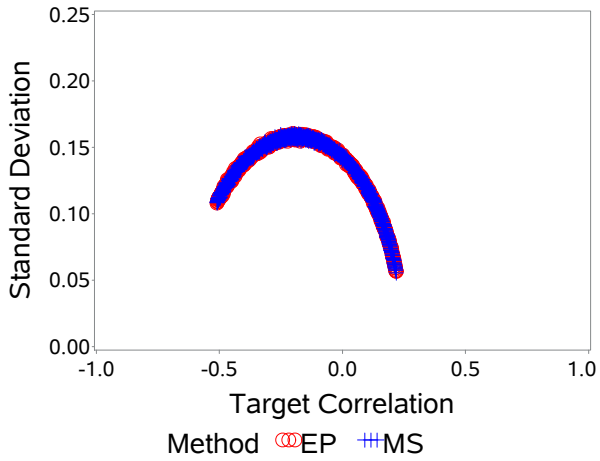
The bias for the MS method was close to zero for all points, with a somewhat parabolic shape, going from 0 to a maximum of 0.014 then decreasing to -0.004. The bias for the EP method was



(a) Proportion of simulations falling within the Fréchet bounds

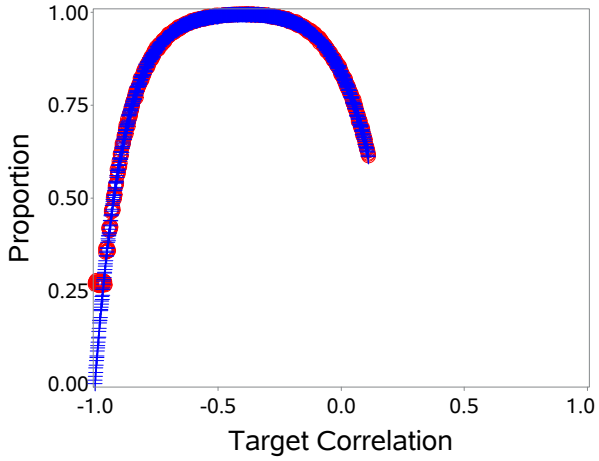


(b) Bias of the estimated correlation

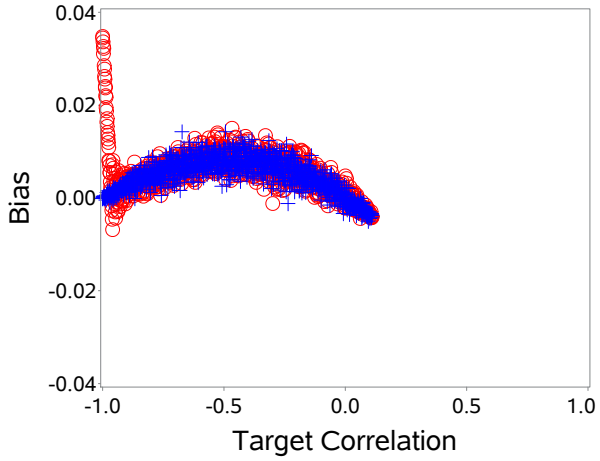


(c) Standard deviation of the estimated correlation

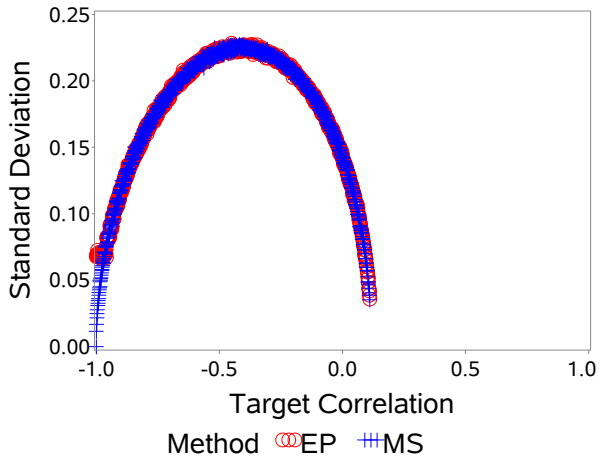
Fig. 2.4. Case: $p_1 = 0.1$, $p_2 = 0.7$



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.5. Case: $p_1 = 0.1$, $p_2 = 0.9$

more extreme at the lower bound of -1, at 0.035, staying near zero with a similar parabolic shape to the MS method otherwise. Being nearly twice the maximum bias of the MS method, this indicates that the EP method does not estimate the correlation near the lower Fréchet bound well.

At the lower bound, the MS method was calculated to have a standard deviation of zero, while the EP method had a standard deviation of 0.067. Both methods increased to 0.228 in the center, then decreased to 0.035 at the upper bound. The discrepancy at the lower bound is due to the inferior estimation of the EP method at the target correlation -1.

Note that this case is symmetric to the case where $p_1 = 0.1$, $p_2 = 0.1$.

Case: $p_1 = 0.3$, $p_2 = 0.3$

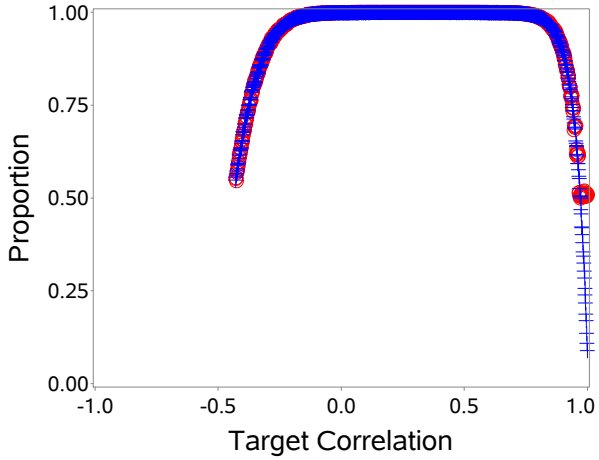
The Fréchet range for this case is [-0.429, 1.000]. The proportion of simulations falling within the Fréchet bounds in the MS case started at 0.543 at the lower bound, increased quickly to 1 and remained there throughout the middle of the range, then dropped to 0.086 at the upper bound. In the EP case, the proportion started at 0.536, increased equally quickly to 1 and remained there throughout the middle of the range, then dropped, leveling off around 0.514.

The bias for the MS method was close to zero for all points, with the most extreme deviation from zero being -0.005. The EP method had much higher maximum bias, with -0.030 at the upper bound, with decreasing bias leading to it in a sharp downward spike. Otherwise the EP method stayed near zero.

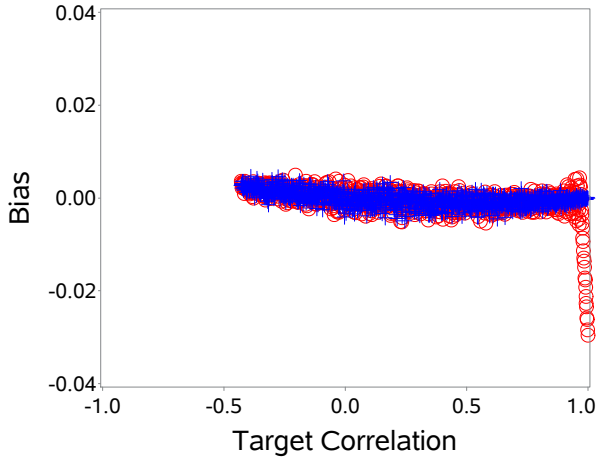
At the lower bound, the standard deviation for both methods was calculated to be 0.072, increasing to 0.150 near target correlation 0.175. The standard deviation of the MS method then decreased to zero at the upper bound, while the EP method only decreased to 0.038.

Case: $p_1 = 0.3$, $p_2 = 0.5$

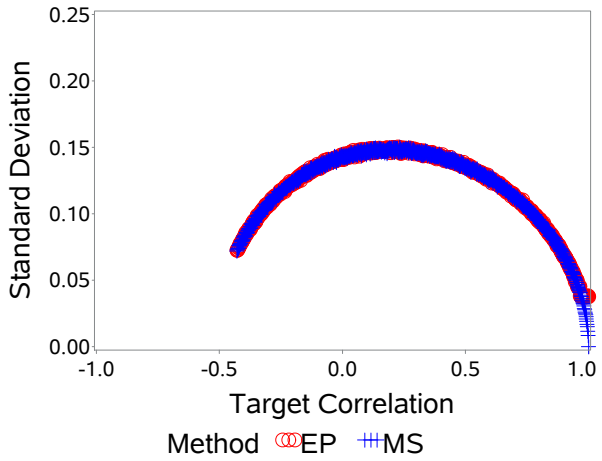
The Fréchet range for this case is [-0.654, 0.654]. The proportion of simulations falling within the Fréchet bounds was nearly the same for both methods. Starting at about 0.492, increasing at similar rates to 1, then decreasing at similar rates to approximately 0.495, producing a parabola



(a) Proportion of simulations falling within the Fréchet bounds

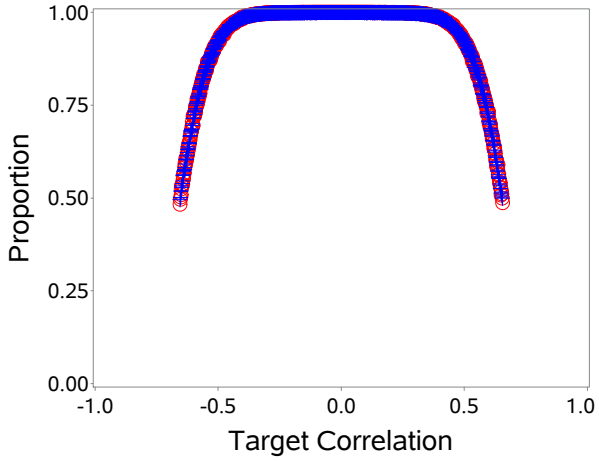


(b) Bias of the estimated correlation

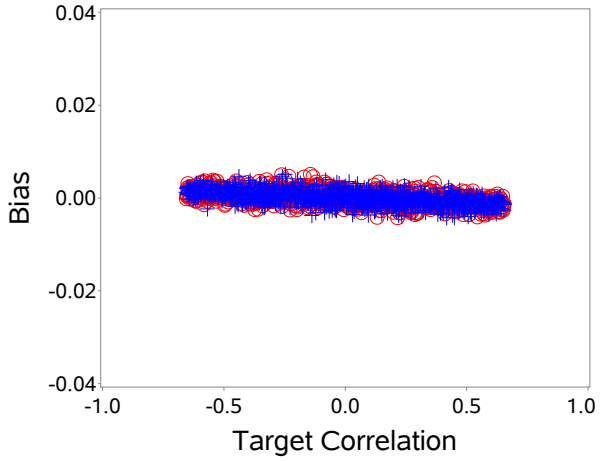


(c) Standard deviation of the estimated correlation

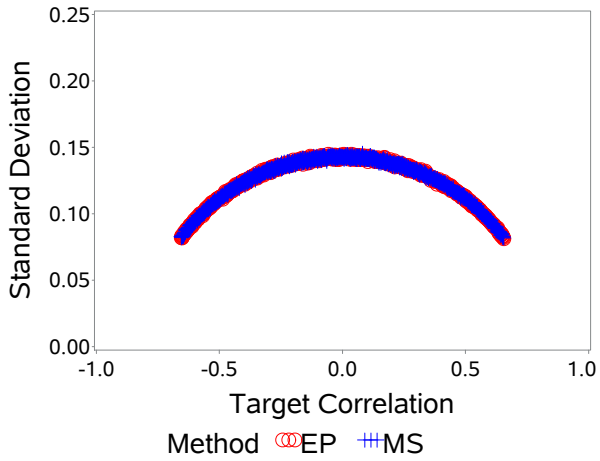
Fig. 2.6. Case: $p_1 = 0.3$, $p_2 = 0.3$



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.7. Case: $p_1 = 0.3$, $p_2 = 0.5$

with a flat peak.

The bias for both methods stayed near zero, with a slight positive slope from the lower bound to the upper bound. The most extreme variation from zero for the MS case was 0.005, and for the EP case it was -0.005.

At the lower bound, the standard deviation for both methods was calculated to be 0.082, increasing to 0.145 near target correlation zero, then returning to 0.083 at the upper bound.

Case: $p_1 = 0.3$, $p_2 = 0.7$

The Fréchet range for this case is [-1.000, 0.429]. The proportion of simulations falling within the Fréchet bounds started at 0.081 for the MS case and at 0.502 in the EP case at the lower bound. The EP case plateaued for a short time until rising at the same rate as the MS case until reaching a maximum of 1, then both decreased to about 0.550 at the upper bound.

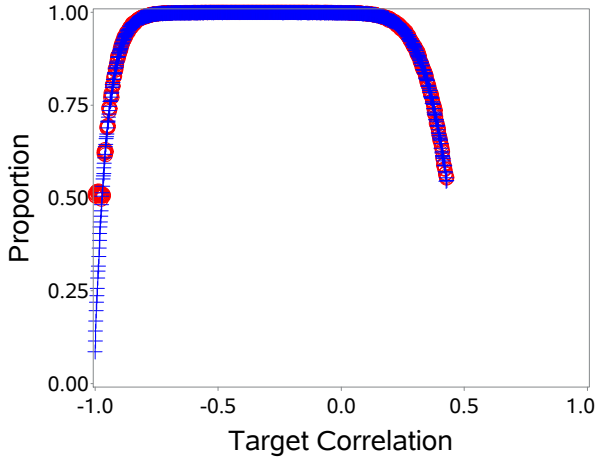
The bias for the MS method was close to zero for all points, with the most extreme deviation from zero being 0.005. The EP method had much higher maximum bias, with 0.029 at the lower bound and a sharp decrease to near zero as the target correlation increased. Otherwise the EP method stayed near zero.

At the lower bound, the MS method was calculated to have a standard deviation of zero, while the EP method had a standard deviation of 0.038. Both methods increased to 0.150 near target correlation -0.205, then decreased to 0.072 at the upper bound.

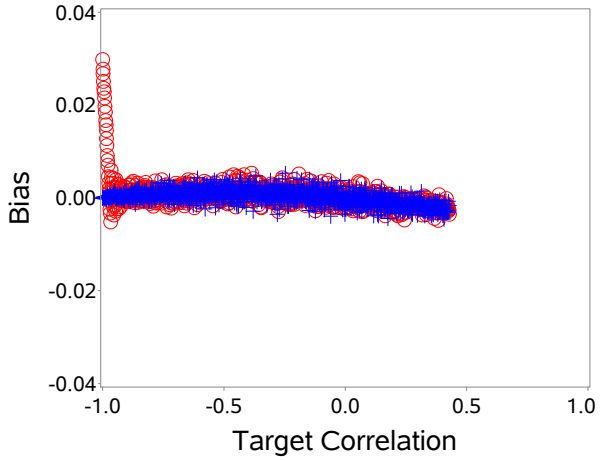
Note that this is symmetric to the case where $p_1 = 0.3$, $p_2 = 0.3$.

Case: $p_1 = 0.3$, $p_2 = 0.9$

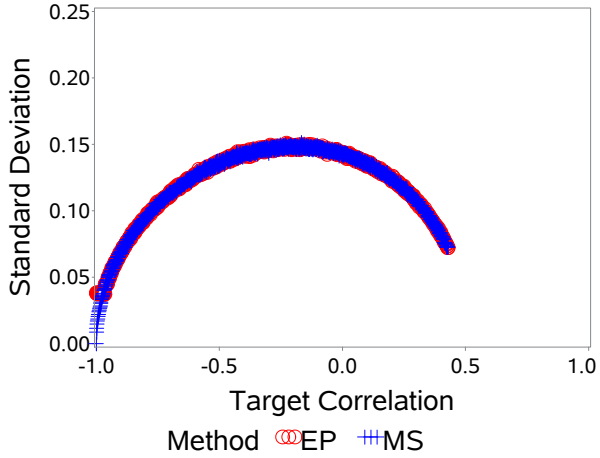
The Fréchet range for this case is [-0.509, 0.218]. The proportion of simulations falling within the Fréchet bounds was nearly identical between the two methods. At the lower bound, the proportion started at 0.527 in the MS case and 0.538 in the EP case. Both increased at about the same rate until reaching over 0.991, then decreased similarly until ending at 0.559 in the MS case and 0.547 in the EP case at the upper bound.



(a) Proportion of simulations falling within the Fréchet bounds

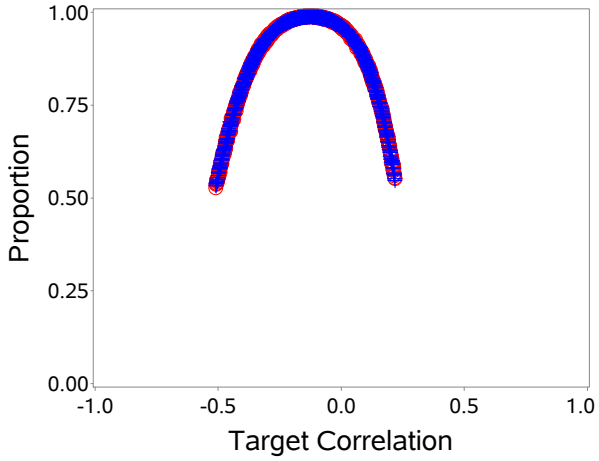


(b) Bias of the estimated correlation

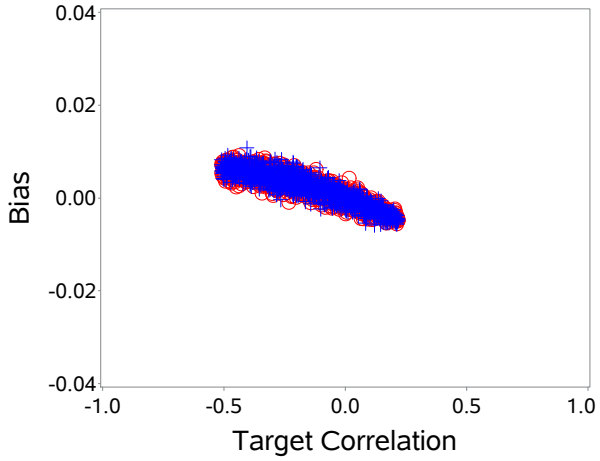


(c) Standard deviation of the estimated correlation

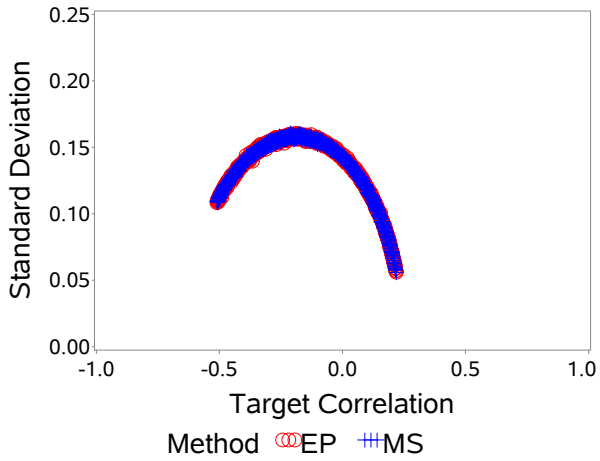
Fig. 2.8. Case: $p_1 = 0.3$, $p_2 = 0.7$



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



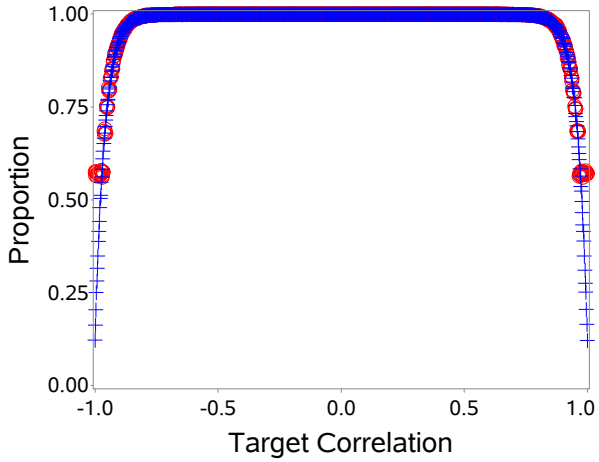
(c) Standard deviation of the estimated correlation

Fig. 2.9. Case: $p_1 = 0.3$, $p_2 = 0.9$

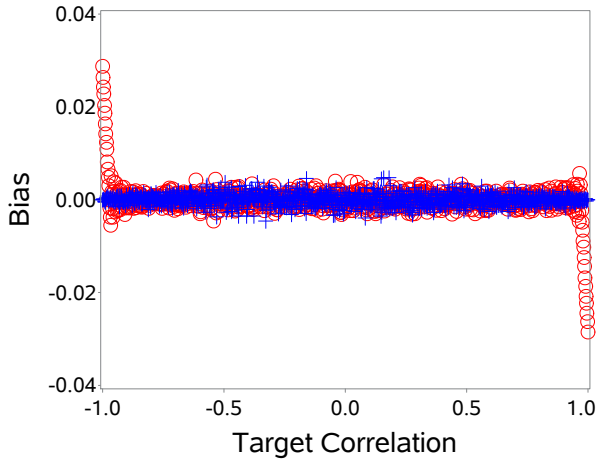
The bias for both methods stayed near zero, with a slight positive slope from the lower bound to the upper bound. The most extreme variation from zero for the both cases was 0.010.

At the lower bound, the standard deviation for both methods was calculated to be 0.109, increasing to 0.160 near target correlation -0.172, then decreasing to 0.056 at the upper bound.

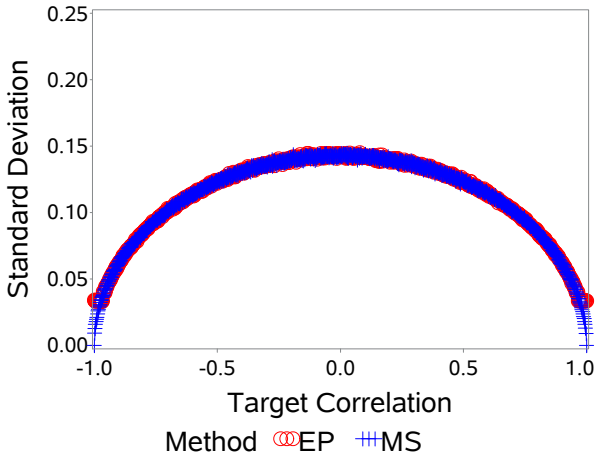
Case: $p_1 = 0.5$, $p_2 = 0.5$



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.10. Case: $p_1 = 0.5$, $p_2 = 0.5$

The Fréchet range for this case is [-1.000, 1.000]. The proportion of simulations falling within

the Fréchet bounds started at 0.118 for the MS case and at 0.567 in the EP case at the lower bound. The EP case plateaued for a short time until rising at the same rate as the MS case until reaching a maximum of 1. In the MS case, the proportion decreased to 0.124 at the upper bound while in the EP case, it decreased to and plateaued around 0.572.

The bias for both methods stayed near zero throughout most of the range. For the EP case, however, comparatively extreme bias was seen at the lower and upper bounds, at 0.028 and -0.028, respectively. The most extreme bias for the MS case was 0.004.

At the lower bound and upper bounds, the MS method was calculated to have a standard deviation of zero, while the EP method had a standard deviation of 0.033. Both methods increased to 0.145 near target correlation zero.

Case: $p_1 = 0.5$, $p_2 = 0.7$

The Fréchet range for this case is [-0.654, 0.654]. The proportion of simulations falling within the Fréchet bounds was nearly the same for both methods. Starting at about 0.495, increasing at similar rates to 1, then decreasing at similar rates to approximately 0.495, producing a parabola with a flat peak.

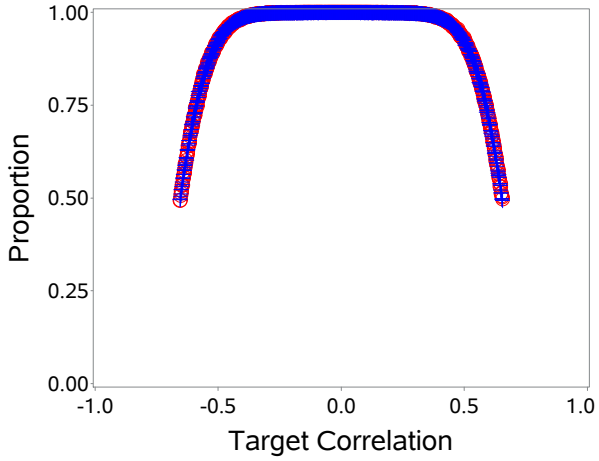
The bias for both methods stayed near zero, with a slight positive slope from the lower bound to the upper bound. The most extreme variation from zero for the both cases was -0.004 for the MS case and 0.004 for the EP case.

At the lower bound, the standard deviation for both methods was calculated to be 0.081, increasing to 0.145 near target correlation 0.065, then decreasing back to 0.082 at the upper bound.

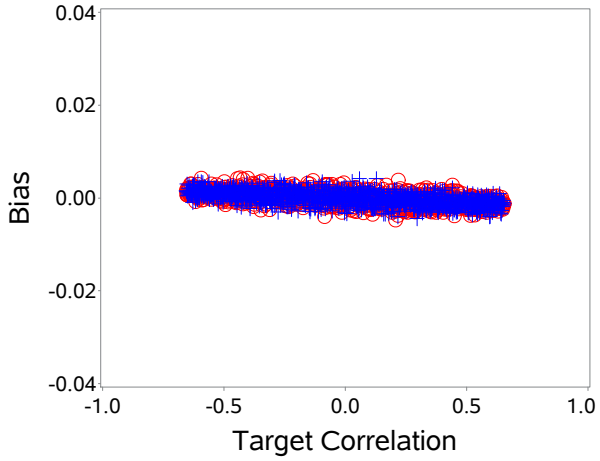
Note that this case is nearly identical to the case where $p_1 = 0.3$, $p_2 = 0.5$.

Case: $p_1 = 0.5$, $p_2 = 0.9$

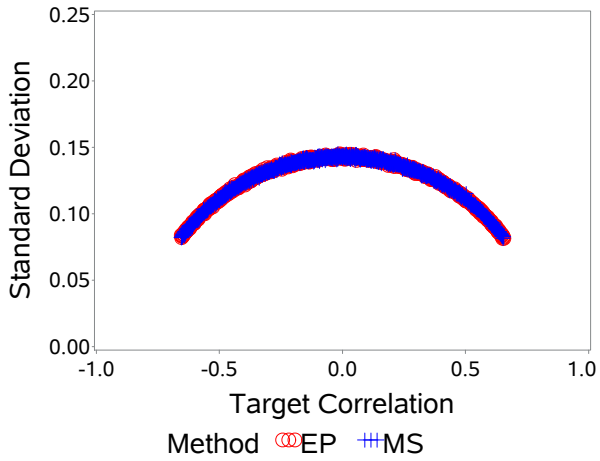
The Fréchet range for this case is [-0.333, 0.333]. The proportion of simulations falling within the Fréchet bounds started at 0.548 for the MS case and at 0.553 in the EP case at the lower bound. Both increased in similar fashion until reaching a maximum of approximately 0.992, then



(a) Proportion of simulations falling within the Fréchet bounds

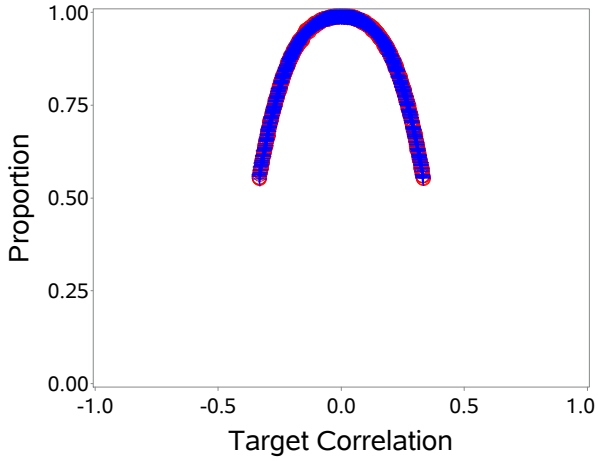


(b) Bias of the estimated correlation

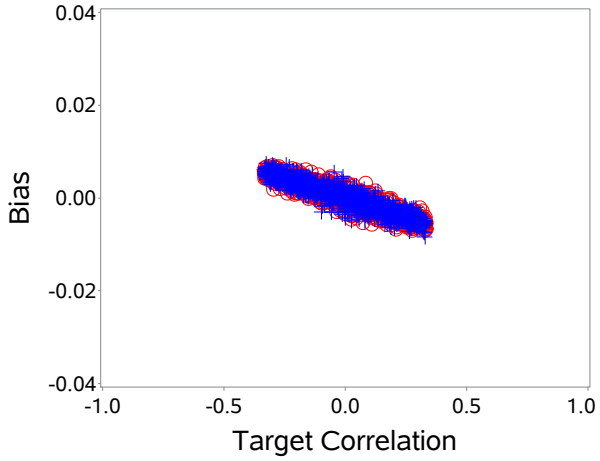


(c) Standard deviation of the estimated correlation

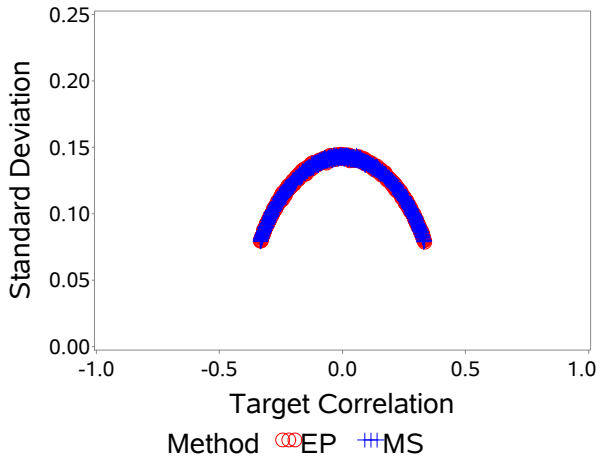
Fig. 2.11. Case: $p_1 = 0.5$, $p_2 = 0.7$



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.12. Case: $p_1 = 0.5$, $p_2 = 0.9$

decreased. The MS case decreased to 0.549 at the upper bound, and the EP case decreased to 0.551.

The bias for both methods stayed near zero, with a slight positive slope from the lower bound to the upper bound. The most extreme variation from zero for the MS case was -0.007 and for the EP case it was 0.007.

At the lower bound, the standard deviation for both methods was calculated to be 0.079, increasing to 0.145 near target correlation zero, then decreasing back to 0.079 at the upper bound.

Note that this case is nearly identical to the case where $p_1 = 0.1$, $p_2 = 0.5$.

Case: $p_1 = 0.7$, $p_2 = 0.7$

The Fréchet range for this case is [-0.429, 1.000]. The proportion of simulations falling within the Fréchet bounds started at 0.551 for the MS case and at 0.544 in the EP case at the lower bound. Both increased in similar fashion until reaching and plateauing at 1. In the MS case, the proportion then decreased to 0.089 at the upper bound while in the EP case, it decreased to and plateaued around 0.514. Note that this case is nearly symmetric to the case $p_1 = 0.3$, $p_2 = 0.3$.

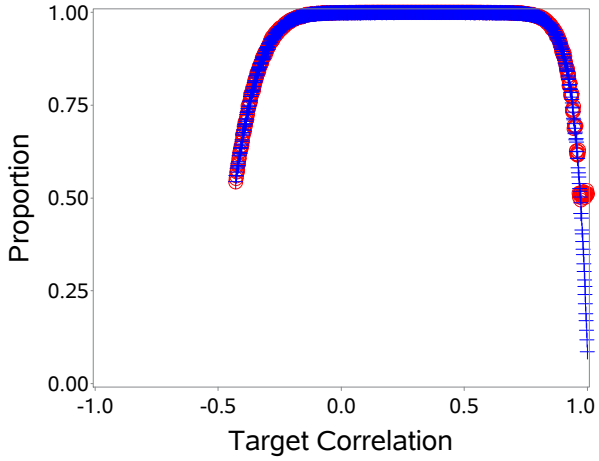
The bias for both methods stayed near zero, except for the EP method at the upper bound, which spiked sharply downward until reaching -0.030. The maximum bias for the MS method was -0.005. There was a slight parabolic pattern to the biases in both cases.

At the lower bound, both methods were calculated to be 0.072, increasing to 0.151 near target correlation 0.180, then decreasing to zero in the MS case at the upper bound of 1, while only decreasing to 0.038 in the EP case, again due to the poor estimation of the target correlation at 1.

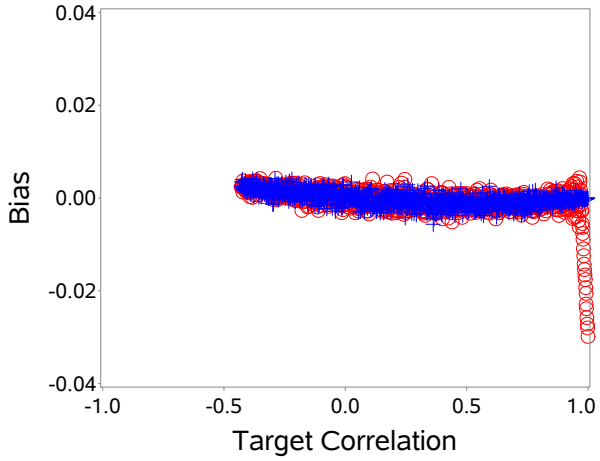
Note that this case is nearly identical to the case where $p_1 = 0.3$, $p_2 = 0.3$ and symmetric to the case where $p_1 = 0.3$, $p_2 = 0.7$.

Case: $p_1 = 0.7$, $p_2 = 0.9$

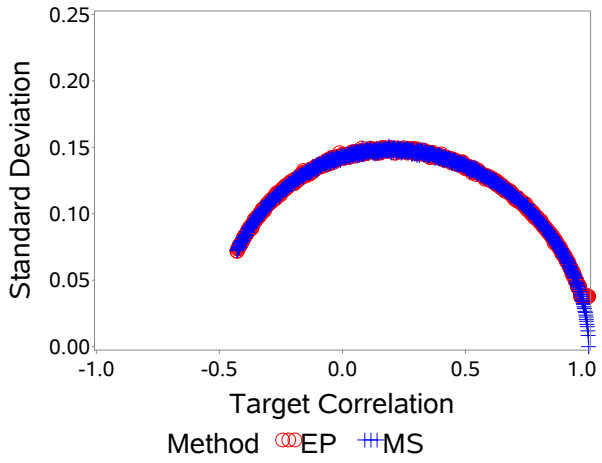
The Fréchet range for this case is [-0.218, 0.509]. The proportion of simulations falling within the Fréchet bounds started at 0.550 for the MS case and at 0.553 in the EP case at the lower



(a) Proportion of simulations falling within the Fréchet bounds

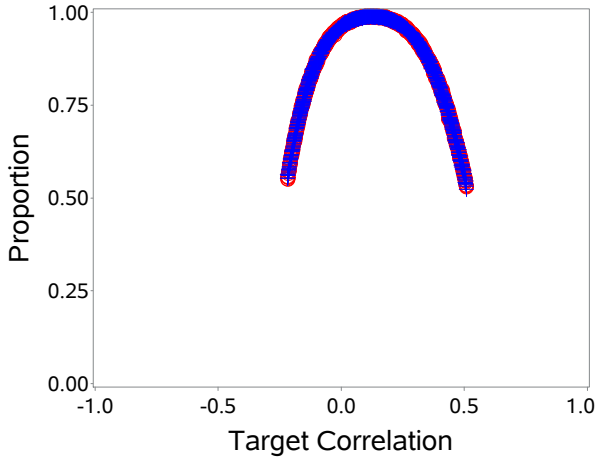


(b) Bias of the estimated correlation

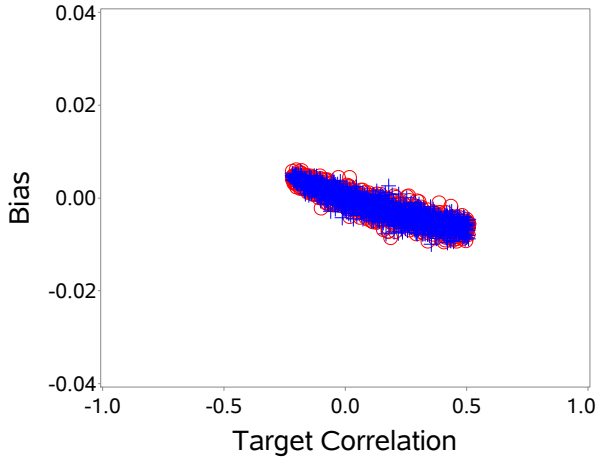


(c) Standard deviation of the estimated correlation

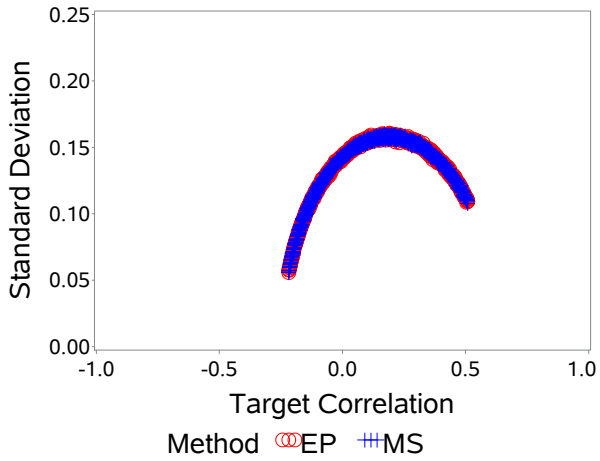
Fig. 2.13. Case: $p_1 = 0.7$, $p_2 = 0.7$



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.14. Case: $p_1 = 0.7$, $p_2 = 0.9$

bound. Both increased in similar fashion until reaching a maximum of approximately 0.992, then decreased to 0.530 at the upper bound.

The bias for both methods stayed near zero, with a slight negative slope from the lower bound to the upper bound. The most extreme variation from zero for the both cases was -0.010.

At the lower bound, the standard deviation for both methods was calculated to be 0.056, increasing to 0.160 near target correlation 0.168, then decreasing to 0.109 at the upper bound.

Note that this case is nearly identical to the case where $p_1 = 0.1$, $p_2 = 0.3$ and symmetric to the case where $p_1 = 0.3$, $p_2 = 0.9$.

Case: $p_1 = 0.9$, $p_2 = 0.9$

The Fréchet range for this case is [-0.111, 1.000]. The proportion of simulations falling within the Fréchet bounds started at 0.615 for the MS case and at 0.609 in the EP case at the lower bound. Both increased in similar fashion until reaching a maximum of approximately 0.997, then decreased. The MS case decreased to 0.162 at the upper bound, and the EP case decreased to and plateaued around 0.379.

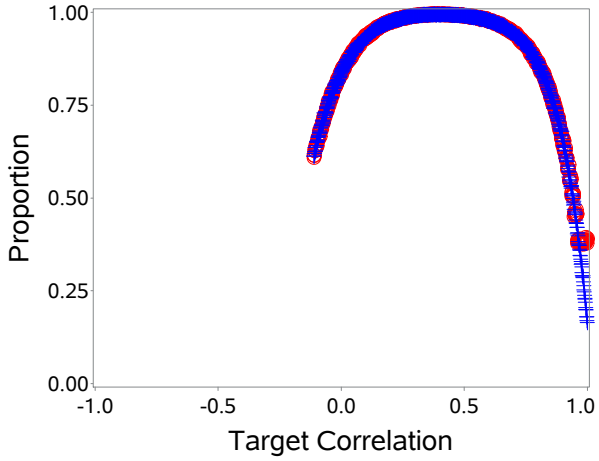
The bias for both methods stayed near zero, except for the EP method at the upper bound, which spiked sharply downward until reaching -0.035. The maximum bias for the MS method was -0.014. There was a slight parabolic pattern to the biases in both cases.

At the lower bound, both methods were calculated to be 0.036, increasing to 0.228 near target correlation 0.430, then decreasing to zero in the MS case at the upper bound of 1, while only decreasing to 0.067.

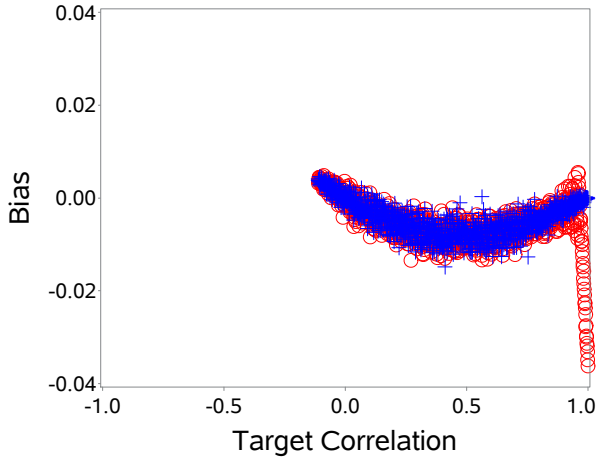
Note that this case is nearly identical to the case where $p_1 = 0.1$, $p_2 = 0.1$ and nearly symmetric to the case where $p_1 = 0.1$, $p_2 = 0.9$.

Remarks Regarding Two-Measure Cases

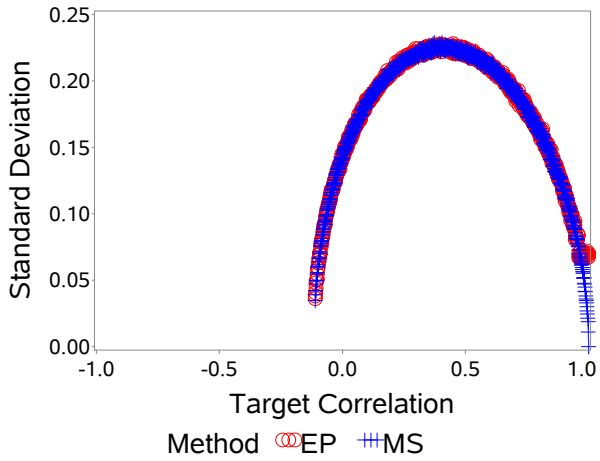
In general the two methods are nearly equivalent in all measures. However, when the correlation being estimated is near 1 or -1, the EP method breaks down and continues to estimate the same



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.15. Case: $p_1 = 0.9$, $p_2 = 0.9$

quantity. In these cases, the MS method appears superior. Also note that some cases have a type of symmetry or are identical. Cases with the same Fréchet bounds on the target correlation have quite similar results. Cases with the same Fréchet bounds, but opposite signs (e.g. [-1.000, 0.429] and [-0.429, 1.000]) have results that are symmetric about zero. From the similarities between certain cases, it appears that the results of the simulations are not dependent upon the values of the marginal probabilities but upon the relative distances from 0.5. This makes intuitive sense due to the nature of the formulae of the bounds (see Section 2.2).

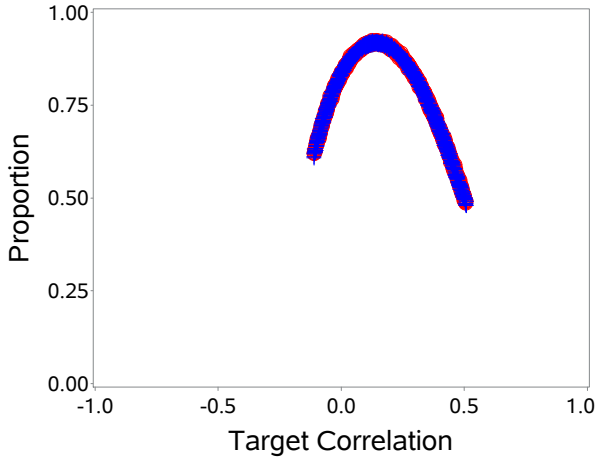
2.4.2.2 Three-Measure Cases

In the three-measure cases, there is redundancy among the cases with a CS correlation matrix, since the order of p_1, p_2, and p_3 should not matter in the calculations. Among the cases with an AR(1) correlation matrix, however, the order of the marginal probabilities matters, so there is less redundancy among the possible cases. In light of the findings from the two-variable cases– the results relying upon the relative distances from 0.5–and the number of possible cases for a three-variable system, only three cases are presented, the same cases are explored for both CS and AR(1). These will be representative of cases that might be seen in an actual study. The CS cases will be presented first, with the AR(1) cases following.

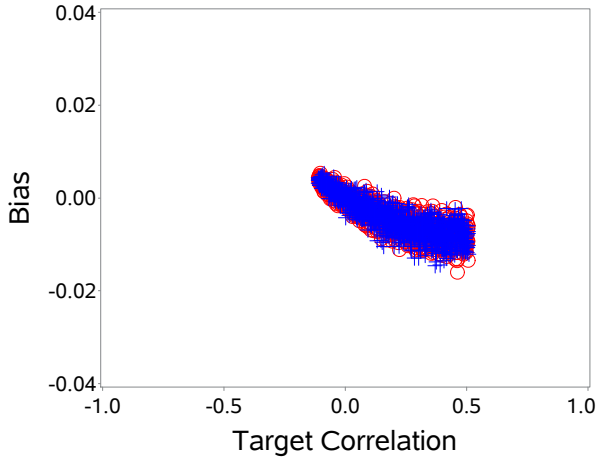
Even though the same quantity ρ is being estimated by $\hat{\rho}_{12}$, $\hat{\rho}_{13}$, and $\hat{\rho}_{23}$ in the CS cases, the process is slightly different for each estimate. Therefore, the statistics for each estimate will be presented separately.

Case (CS): $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.3$

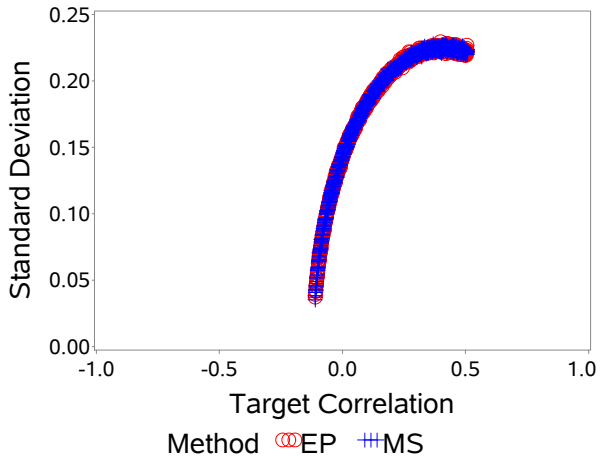
Representing a case where the marginal means stay steady then increase, this case has Fréchet bounds [-0.111,0.509]. However, due to the need for the correlation matrix to be invertible while using the EP method, the method cannot be used at the endpoints. Therefore, the EP method has estimable range [-0.110, 0.507] as seen in the simulations. The MS method has the advantage of being estimable over the entire Fréchet range.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.16. Case (CS): $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.3$. Figures for $\hat{\rho}_{12}$.

For $\hat{\rho}_{12}$, the proportion of simulations with a mean estimated correlation falling within the Fréchet bounds in the EP case started at 0.620, increased to a maximum of 0.925 at target correlation 0.142, then dropped to 0.491 at the EP upper bound. In the MS case, the proportion started at the lower Fréchet bound at 0.611, increased to 0.924 at target correlation 0.126, then decreased to 0.490 at the upper Fréchet bound.

The bias of the estimated correlation had a "fan" shape, with the maximum bias of -0.016 for the EP method occurring at target correlation 0.463, and the maximum bias of -0.015 for the MS method occurring at 0.371.

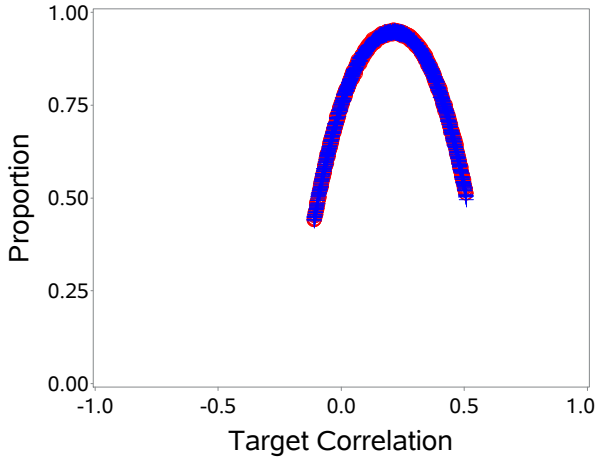
The standard deviation of the estimated correlation in the EP case started at 0.037, increasing to a maximum of 0.230 at target correlation 0.397, then decreasing slightly to 0.222 at the EP upper bound. In the MS case, the standard deviation started at 0.035 at the lower Fréchet bound, increased to a maximum of 0.228 at target correlation 0.420, then decreased to 0.223.

For $\hat{\rho}_{13}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.441, increased to a maximum of 0.952 at target correlation 0.214, then dropped to 0.512 at the EP upper bound. In the MS case, the proportion started at the lower Fréchet bound at 0.448, increased to 0.952 at target correlation 0.219, then decreased to 0.495 at the upper Fréchet bound.

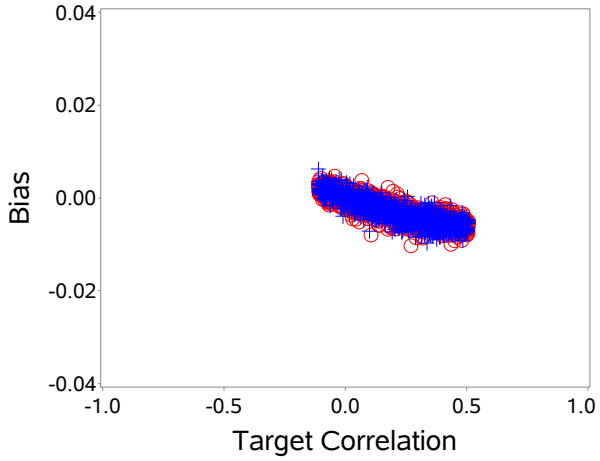
The bias had a decreasing slope, with the maximum bias of -0.010 for the EP method occurring at target correlation 0.271. For the MS method, the maximum of -0.010 occurred at target correlation 0.337.

The standard deviation in the EP case started at 0.115, increasing to a maximum of 0.160 at target correlation 0.181, then decreasing to 0.107 at the EP upper bound. In the MS case, the standard deviation started at the lower Fréchet bound at 0.116, increased to a maximum of 0.161 at target correlation 0.156, then decreased to 0.109 at the upper Fréchet bound.

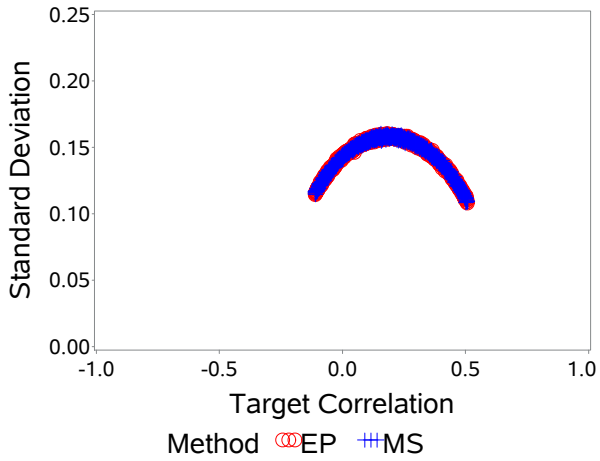
For $\hat{\rho}_{23}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.444, increased to a maximum of 0.953 at target correlation 0.214, then dropped to 0.514 at the EP upper bound. For the MS case, the proportion started at 0.435 at the lower Fréchet bound, increased to 0.954 at target correlation 0.204, then decreased to 0.109 at the upper Fréchet bound.



(a) Proportion of simulations falling within the Fréchet bounds

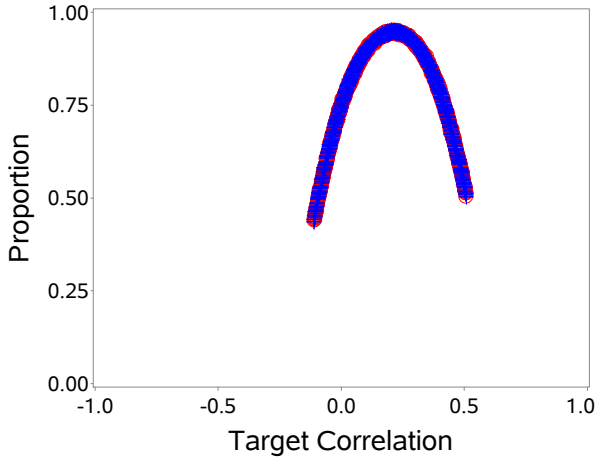


(b) Bias of the estimated correlation

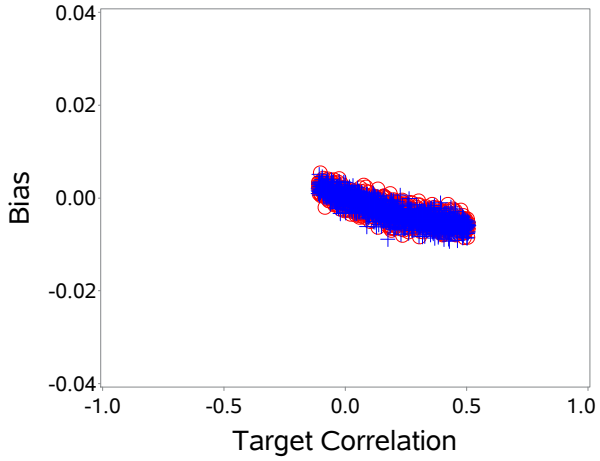


(c) Standard deviation of the estimated correlation

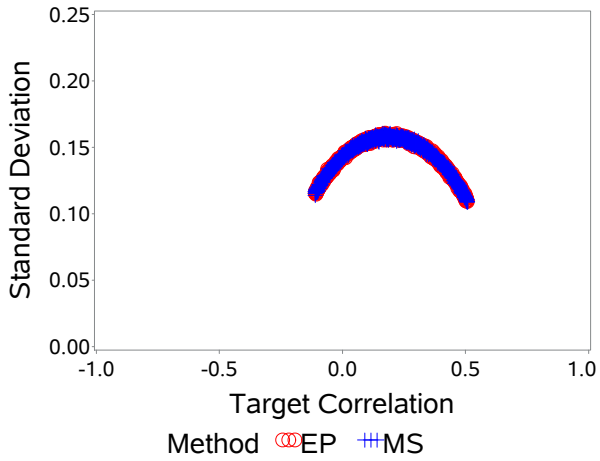
Fig. 2.17. Case (CS): $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.3$. Figures for $\hat{\rho}_{13}$.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.18. Case (CS): $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.3$. Figures for $\hat{\rho}_{23}$.

The bias had a slightly decreasing slope, with the maximum bias of -0.009 for the EP method occurring at target correlation 0.491. The maximum for the MS method was also -0.009, occurring at target correlation 0.429.

The standard deviation in the EP case started at 0.116, increasing to a maximum of 0.161 at target correlation 0.173, then decreasing to 0.109 at the EP upper bound. In the MS case, the standard deviation started at 0.117 at the lower Fréchet bound, increasing to 0.161 at target correlation 0.164, then decreasing to 0.109 at the upper Fréchet bound.

Case (CS): $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.1$

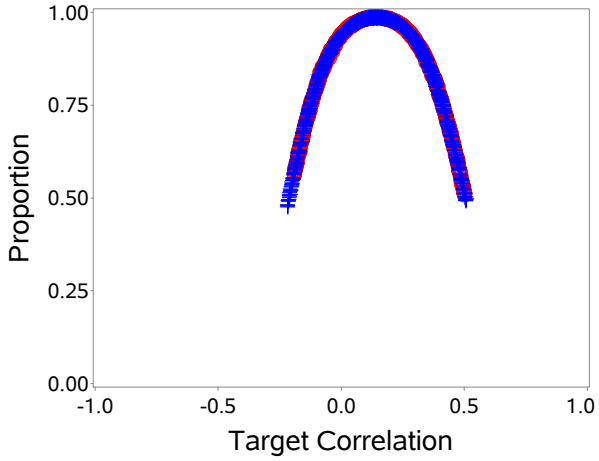
Representing a case where the marginal means stay steady then decrease, this case has Fréchet bounds [-0.218, 0.509]. Again, the EP method cannot be estimated at these endpoints, and the estimable range is [-0.191, 0.503].

For $\hat{\rho}_{12}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.563, increased to a maximum of 0.989 at target correlation 0.142, then dropped to 0.512 at the EP upper bound. In the MS case, this proportion started at 0.482 at the lower Fréchet bound, increased to 0.924 at target correlation 0.126, then decreased to 0.495 at the upper Fréchet bound.

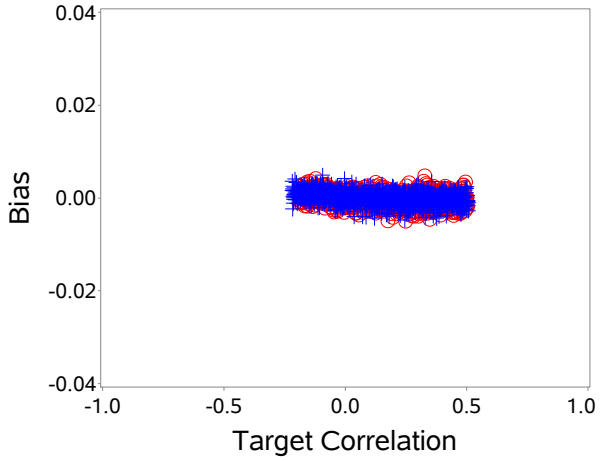
The bias stayed around zero for both methods with a slight negative incline, with the maximum bias of -0.005 for the EP method occurring at target correlation 0.249. For the MS method, the maximum bias of 0.004 occurred at target correlation -0.092.

The standard deviation in the EP case started at 0.124, increasing to a maximum of 0.150 at target correlation 0.218, then decreasing slightly to 0.136 at the EP upper bound. In the MS case, the standard deviation started at 0.121 at the lower Fréchet bound, increased to 0.150 at target correlation 0.179, then decreased to 0.135 at the upper Fréchet bound.

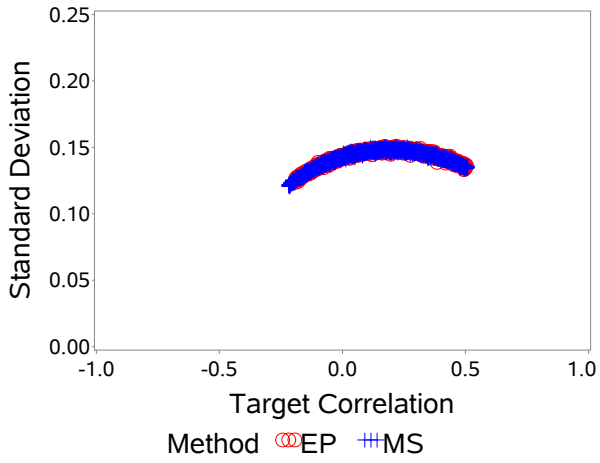
For $\hat{\rho}_{13}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.633, increased to a maximum of 0.991 at target correlation 0.140, then dropped to 0.526 at the EP upper bound. In the MS case, the proportion started at 0.521 at the lower Fréchet bound, increased to 0.952 at target correlation 0.129, then dropped to 0.497 at the upper Fréchet bound.



(a) Proportion of simulations falling within the Fréchet bounds

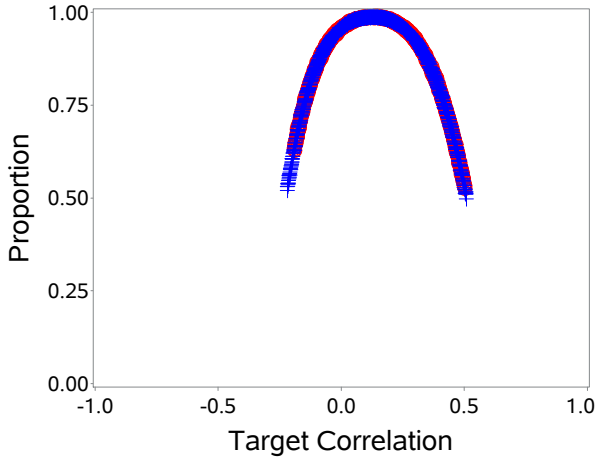


(b) Bias of the estimated correlation

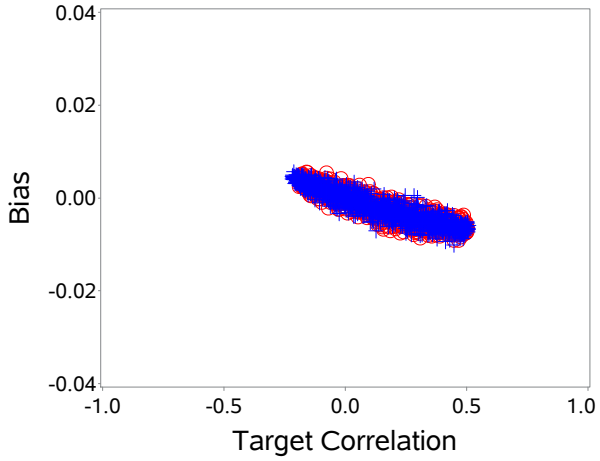


(c) Standard deviation of the estimated correlation

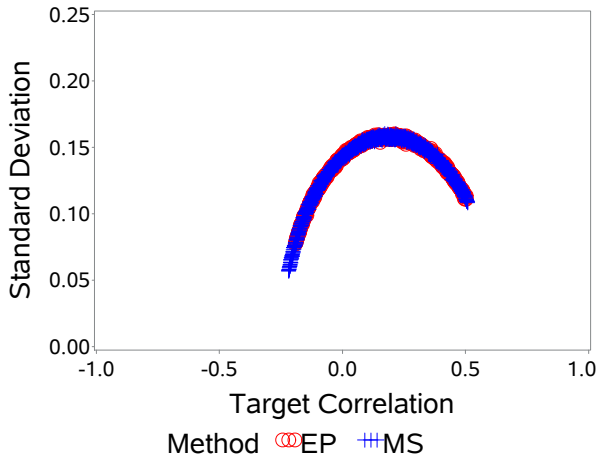
Fig. 2.19. Case (CS): $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.1$. Figures for $\hat{\rho}_{12}$.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation

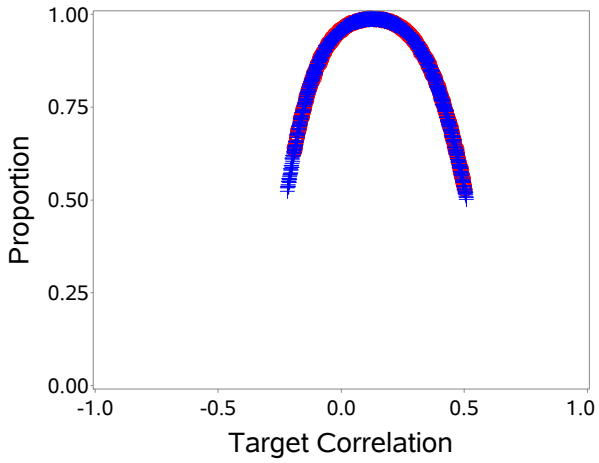


(c) Standard deviation of the estimated correlation

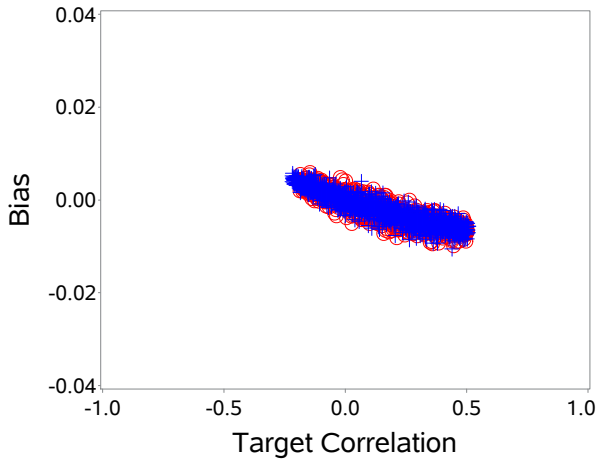
Fig. 2.20. Case (CS): $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.1$. Figures for $\hat{\rho}_{13}$.

The bias stayed around zero for both methods with a slightly negative slope, with the maximum bias of -0.009 for the EP method occurring at target correlation 0.447, and the maximum bias of -0.010 occurring at target correlation 0.449 for the MS method.

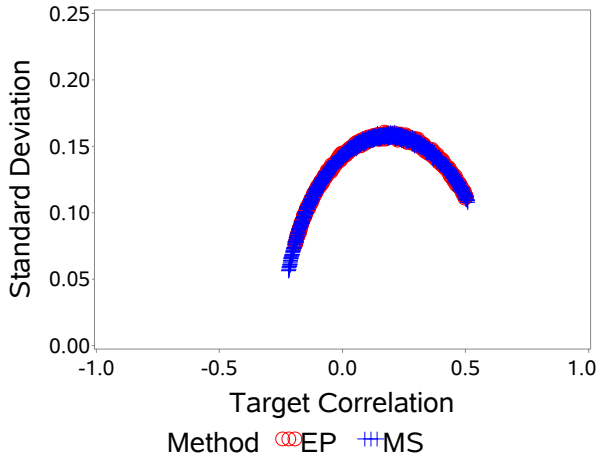
The standard deviation in the EP case started at 0.077, increasing to a maximum of 0.160 at target correlation 0.211, then decreasing to 0.111 at the EP upper bound. For the MS case, the standard deviation started at 0.057 at the lower Fréchet bound, increasing to 0.160 at target correlation 0.172, then decreasing to 0.109 at the upper Fréchet bound.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.21. Case (CS): $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.1$. Figures for $\hat{\rho}_{23}$.

For $\hat{\rho}_{23}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.641, increased to a maximum of 0.991 at target correlation 0.117, then dropped to 0.529 at the EP upper bound. For the MS method, the proportion started at 0.533 at the lower Fréchet bound, increased to 0.991 at target correlation 0.120, then decreased to 0.502 at the upper Fréchet bound.

The bias had a decreasing slope, with the maximum bias of -0.010 for the EP method occurring at target correlation 0.446. The maximum bias for the MS method was -0.011, occurring at target correlation 0.441.

The standard deviation in the EP case started at 0.076, increasing to a maximum of 0.160 at target correlation 0.170, then decreasing to 0.112 at the EP upper bound. In the MS case, the standard deviation started at 0.057 at the lower Fréchet bound, increased to a maximum of 0.161 at target correlation 0.213, then decreased to 0.109 at the upper Fréchet bound.

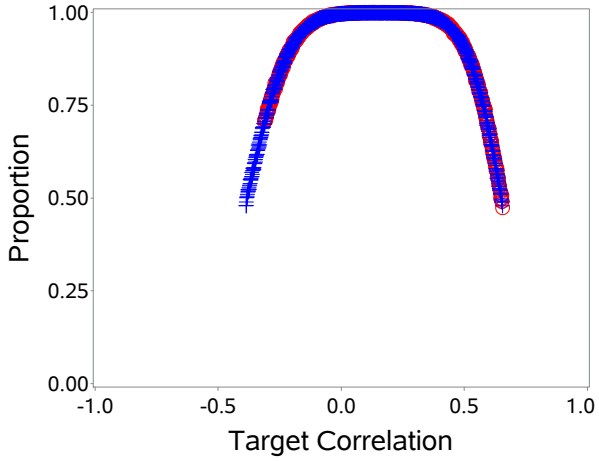
Case (CS): $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.3$

Representing a case where the marginal means steadily decrease, this case has Fréchet bounds [-0.386, 0.654]. The EP method cannot be estimated at the lower bound in this case, so the estimable range is [-0.311, 0.654].

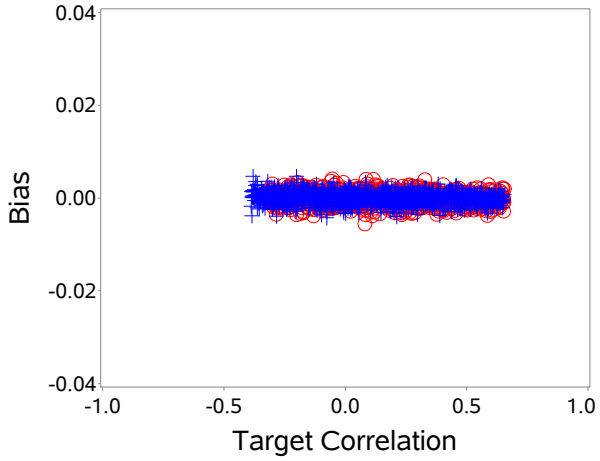
For $\hat{\rho}_{12}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.705 at the lower EP bound, increased to a maximum of 1 throughout the middle of the range, then dropped to 0.474 at the upper Fréchet bound. For the MS case, the proportion started at 0.479 at the lower Fréchet bound, increased to 1 throughout the middle of the range, then dropped back to 0.479 at the upper bound.

The bias stayed close to zero, with the maximum bias of -0.006 for the EP method occurring at target correlation 0.081. For the MS method, the largest bias was 0.005, occurring at target correlation -0.379.

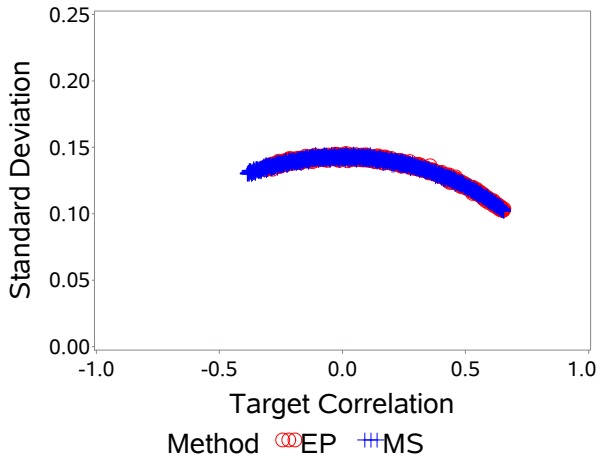
The standard deviation in the EP case started at 0.133, increasing to a maximum of 0.145 at target correlation 0.014, then decreasing to 0.104 at the upper bound. In the MS case, the standard deviation started at 0.130 at the lower Fréchet bound, increased slightly to 0.145 at target



(a) Proportion of simulations falling within the Fréchet bounds



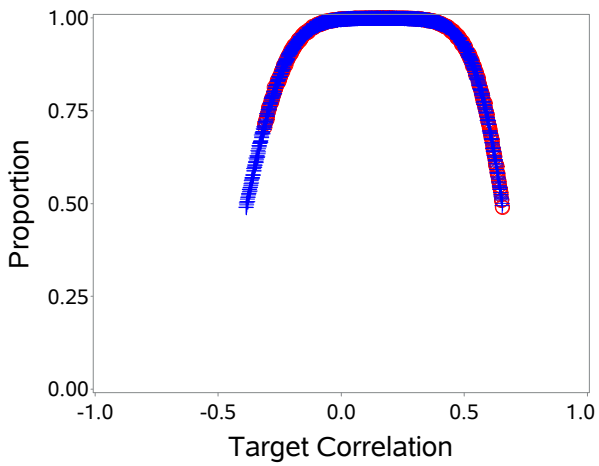
(b) Bias of the estimated correlation



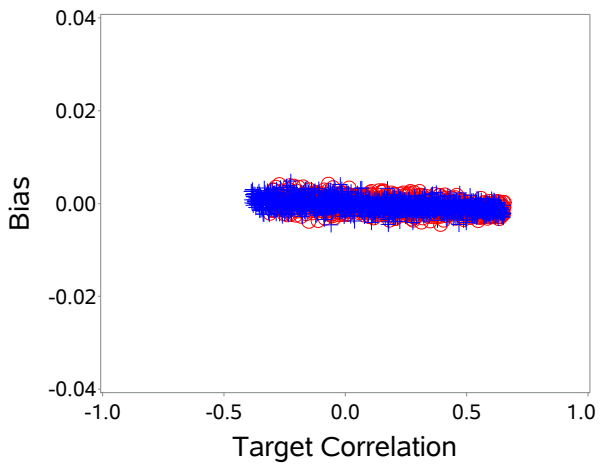
(c) Standard deviation of the estimated correlation

Fig. 2.22. Case (CS): $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.3$. Figures for $\hat{\rho}_{12}$.

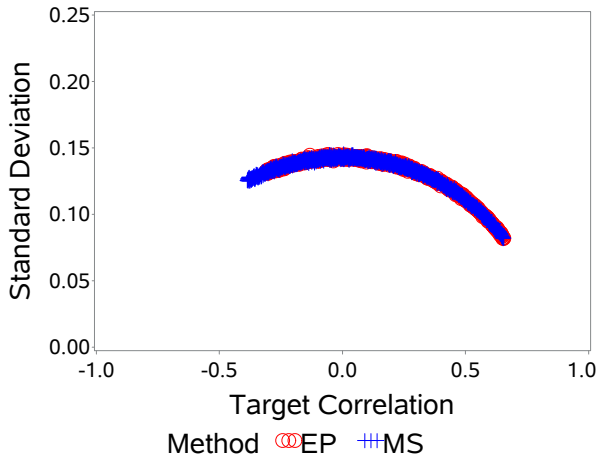
correlation 0.045, then decreased to 0.102 at the upper bound.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

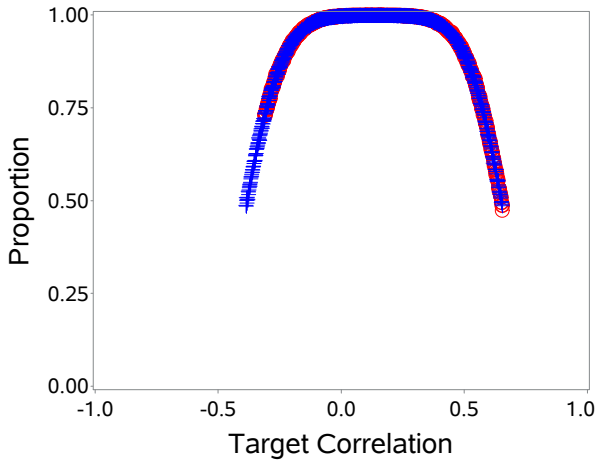
Fig. 2.23. Case (CS): $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.3$. Figures for $\hat{\rho}_{13}$.

For $\hat{\rho}_{13}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.713, increased to a maximum of 1 throughout the middle of the range, then dropped to 0.489 at the upper bound. In the MS case, the proportion started at 0.491 at the lower Fréchet bound, increased to 1 throughout the middle of the range, then decreased to 0.501 at the upper bound.

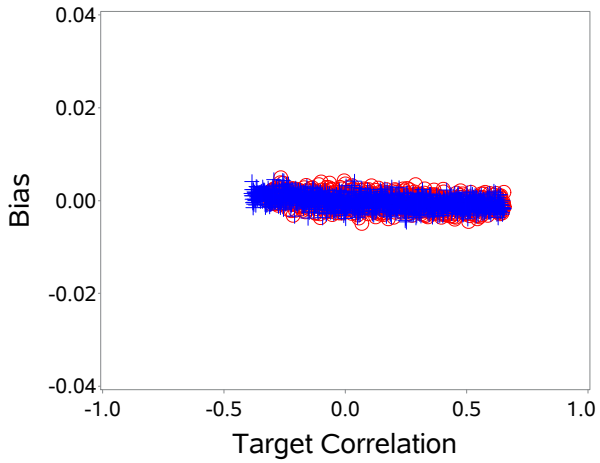
The bias stayed close to zero, with the maximum bias of -0.004 for the EP method occurring at target correlation 0.393. The largest bias for the MS method was 0.005, occurring at target

correlation -0.224.

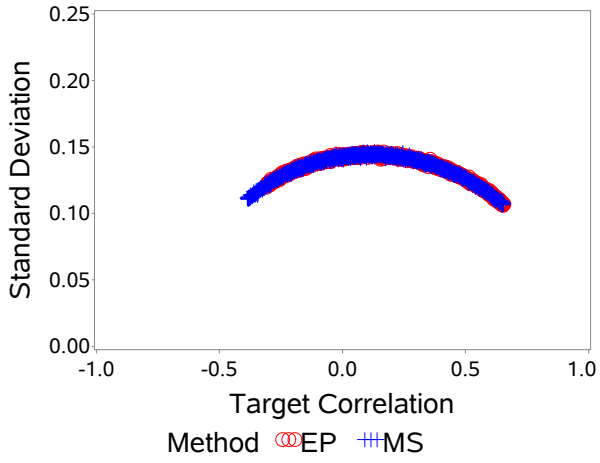
The standard deviation in the EP case started at 0.132, increasing to a maximum of 0.145 at target correlation -0.019, then decreasing to 0.082 at the upper bound. In the MS case, the standard deviation started at 0.126 at the lower Fréchet bound, increased to 0.146 at target correlation 0.004, then decreased to 0.082 at the upper bound.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.24. Case (CS): $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.3$. Figures for $\hat{\rho}_{23}$.

For $\hat{\rho}_{23}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.731, increased to a maximum of 1 throughout the middle of the range, then dropped to 0.487

at the upper bound. In the MS case, the proportion started at 0.487 at the lower Fréchet bound, increased to 1 throughout the middle of the range, then decreased to 0.486 at the upper bound.

The bias stayed close to zero, with the greatest bias of 0.004 for the EP method occurring at target correlation -0.266. The greatest bias for the MS method was -0.005, occurring at target correlation 0.251.

The standard deviation in the EP case started at 0.120, increasing to a maximum of 0.146 at target correlation 0.144, then decreasing to 0.106 at the upper bound. In the MS case, the standard deviation started at 0.111 at the lower Fréchet bound, increased to 0.147 at target correlation 0.129, then decreased to 0.107 at the upper bound.

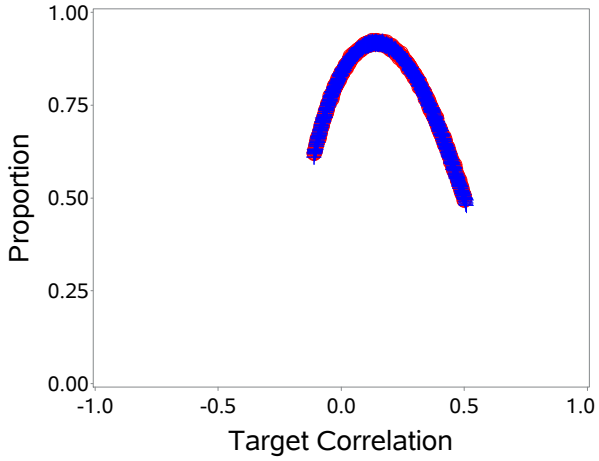
Case (AR(1)): $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.3$

Representing a case where the marginal means stay steady then increase with an AR(1) correlation structure, this case has Fréchet bounds [-0.111, 0.509]. In the EP case, the target correlations near the bounds are inestimable, so the estimable range is [-0.110, 0.501].

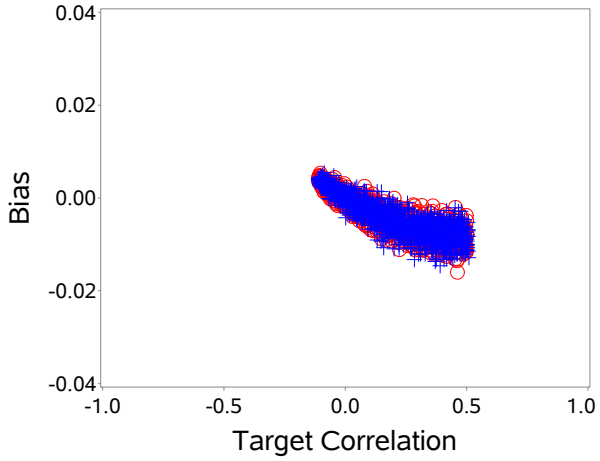
For $\hat{\rho}_{12}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.619 at the lower EP bound, increased to a maximum of 0.925 at target correlation 0.142, then dropped to 0.501 at the upper EP bound. For the MS case, the proportion started at 0.611 at the lower Fréchet bound, increased to a maximum of 0.924 at target correlation 0.169, then decreased back to 0.494 at the upper Fréchet bound.

The bias for both methods had a decreasing slope with a "fan" shape. The greatest bias for the EP method was -0.016, occurring at target correlation 0.463, and for the MS method, it was -0.015, occurring at target correlation 0.391.

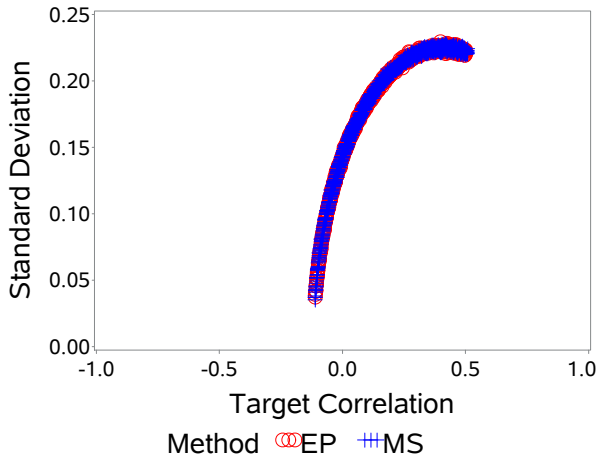
The standard deviation for the estimated correlation under the EP method started at the EP lower bound at 0.037, increased to a maximum of 0.230 at target correlation 0.397, then decreased to 0.221 at the EP upper bound. In the MS case, the standard deviation started at 0.035 at the lower Fréchet bound, increased to 0.228 at target correlation 0.420, then decreased to 0.224 at the upper Fréchet bound.



(a) Proportion of simulations falling within the Fréchet bounds

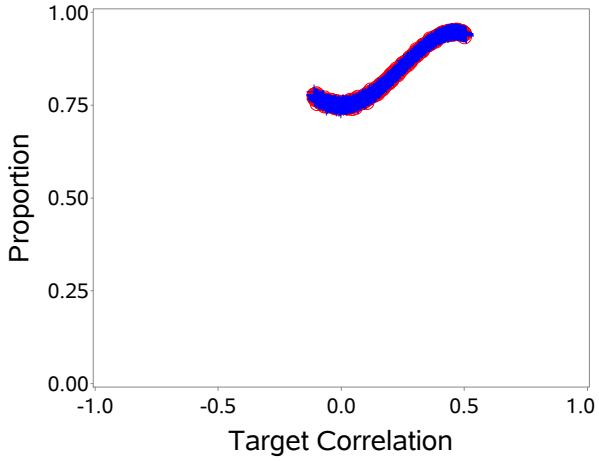


(b) Bias of the estimated correlation

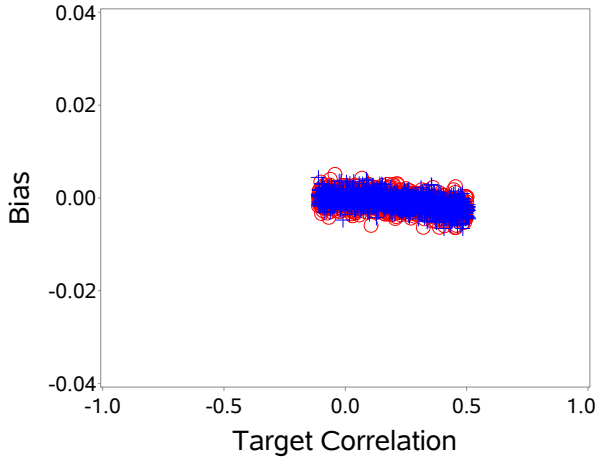


(c) Standard deviation of the estimated correlation

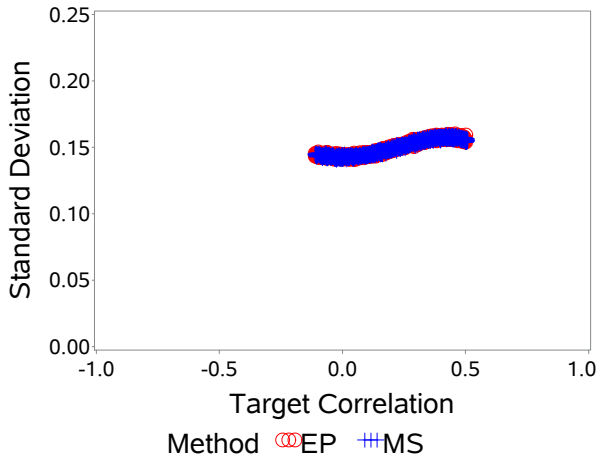
Fig. 2.25. Case (AR(1)): $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.3$. Figures for $\hat{\rho}_{12}$.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.26. Case (AR(1)): $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.3$. Figures for $\hat{\rho}_{13}$.

For $\hat{\rho}_{13}$, the quantity being estimated is ρ^2. For the sake of completion, estimates are presented in reference to ρ. As such, the plots for the proportion and the standard deviation appear as "waves" rather than as partial parabolas.

The proportion of simulations falling within the Fréchet bounds in the EP case started at 0.778, decreased slightly, then increased to a maximum of 0.953 at target correlation 0.468 (ρ^2 = 0.219), then decreased to 0.943. In the MS case, the proportion started at 0.787, decreased slightly, increased to 0.951 at target correlation 0.453 (ρ^2 = 0.205), then decreased to 0.937 at the upper bound.

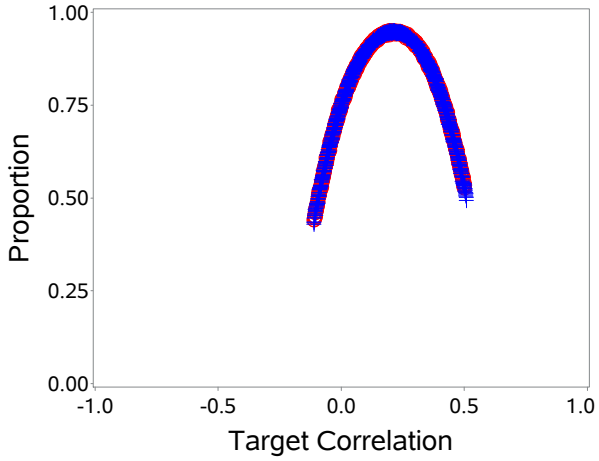
The bias had a slight decreasing slope. For the EP case, the greatest bias was -0.006, which occurred at target correlation 0.453 (ρ^2 = 0.205). In the MS case, the greatest bias was -0.007, occurring at target correlation 0.484 (ρ^2 = 0.234).

The standard deviation for the estimated correlation under the EP method started at 0.143, decreased slightly, then increased to a maximum of 0.160 at target correlation 0.458 (ρ^2 = 0.209), then decreased to 0.154 at the upper EP bound. In the MS case, the standard deviation started at 0.144 at the lower bound, decreased slightly, then increased to a maximum of 0.160 at target correlation 0.421 (ρ^2 = 0.177), then decreased to 0.156 at the upper bound.

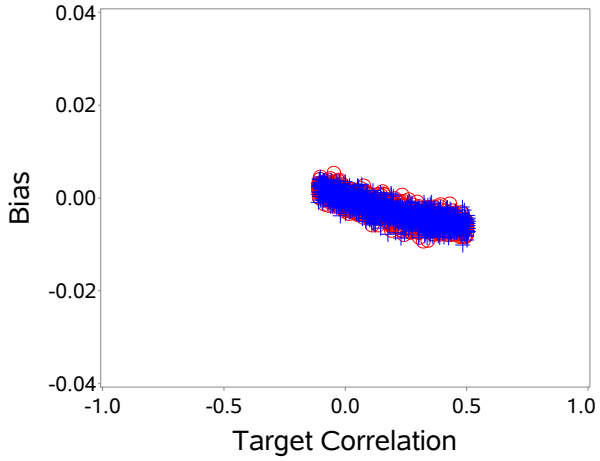
For $\hat{\rho}_{23}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.442 at the lower EP bound, increased to a maximum of 0.951 at target correlation 0.199, then dropped to 0.527 at the upper EP bound. For the MS case, the proportion started at 0.434 at the lower Fréchet bound, increased to a maximum of 0.952 at target correlation 0.195, then decreased back to 0.493 at the upper Fréchet bound.

The bias for both methods had a slightly decreasing slope. The greatest bias for the EP method was -0.009, occurring at target correlation 0.321, and for the MS method, it was -0.010, occurring at target correlation 0.484.

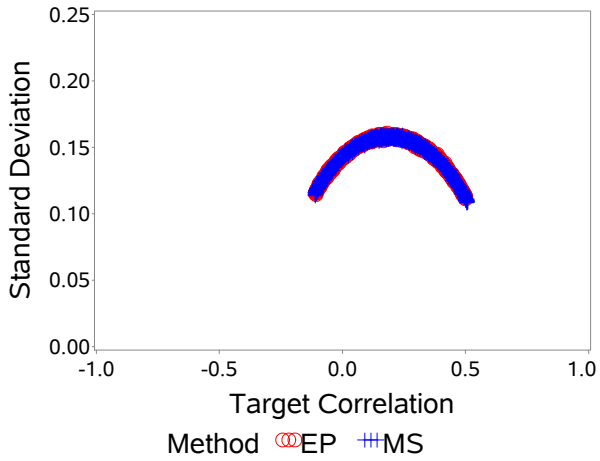
The standard deviation for the estimated correlation under the EP method started at the EP lower bound at 0.116, increased to a maximum of 0.161 at target correlation 0.182, then decreased to 0.111 at the EP upper bound. In the MS case, the standard deviation started at 0.114 at the lower



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



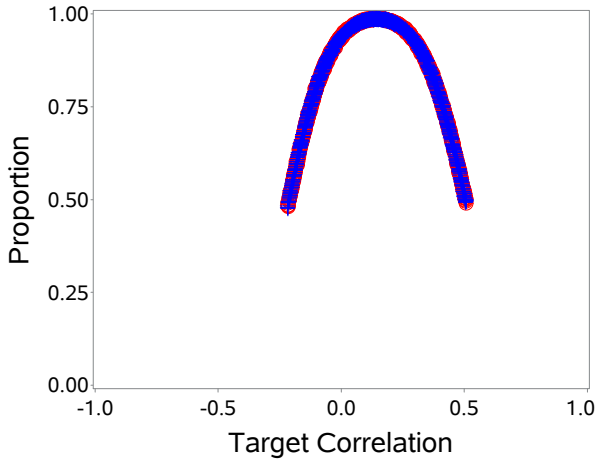
(c) Standard deviation of the estimated correlation

Fig. 2.27. Case (AR(1)): $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.3$. Figures for $\hat{\rho}_{23}$.

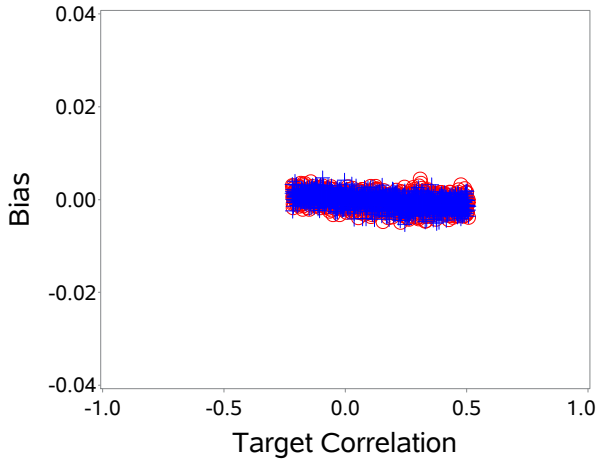
Fréchet bound, increased to 0.160 at target correlation 0.164, then decreased to 0.109 at the upper Fréchet bound.

Case (AR(1)): $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.1$

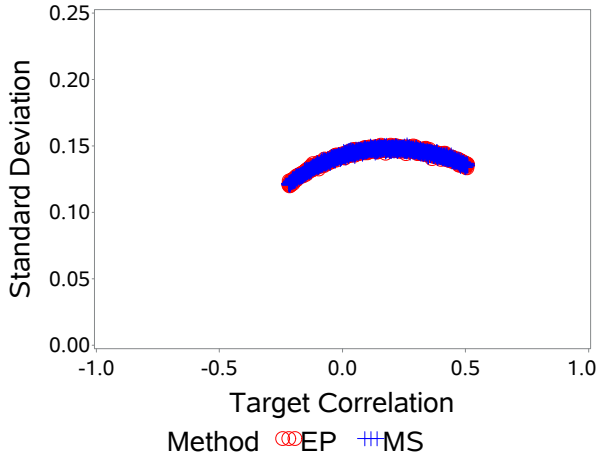
Representing a case where the marginal means stay steady then decrease, this case has Fréchet bounds [-0.218, 0.509]. Again, the EP method is unable to estimate the correlation near the end-points. The estimable range is [-0.217, 0.508].



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.28. Case (AR(1)): $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.1$. Figures for $\hat{\rho}_{12}$.

For $\hat{\rho}_{12}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.491 at the lower EP bound, increased to a maximum of 0.989 at target correlation 0.139, then dropped to 0.502 at the upper EP bound. For the MS case, the proportion started at 0.483 at the lower Fréchet bound, increased to a maximum of 0.989 at target correlation 0.141, then decreased back to 0.497 at the upper Fréchet bound.

The bias for both methods had a somewhat decreasing slope. The greatest bias for the EP method was -0.005, occurring at target correlation 0.229, and for the MS method, it was also -0.005, occurring at target correlation 0.245.

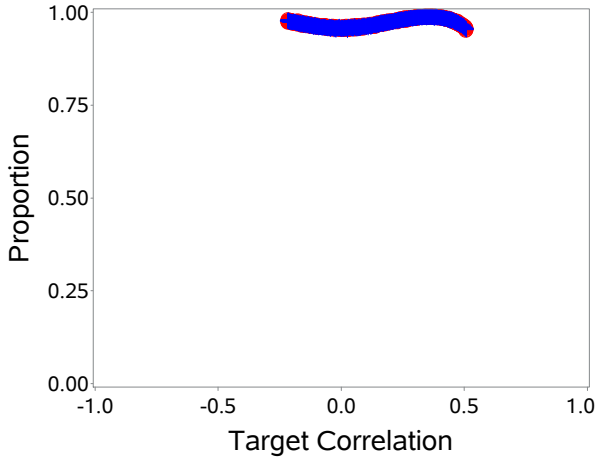
The standard deviation for the estimated correlation under the EP method started at the EP lower bound at 0.121, increased to a maximum of 0.151 at target correlation 0.195, then decreased to 0.136 at the EP upper bound. In the MS case, the standard deviation started at 0.122 at the lower Fréchet bound, increased to 0.151 at target correlation 0.262, then decreased to 0.135 at the upper Fréchet bound.

For $\hat{\rho}_{13}$, the quantity being estimated is ρ^2. For the sake of completion, estimates are presented in reference to ρ. As such, the plots for the proportion and the standard deviation appear as "waves" rather than as partial parabolas.

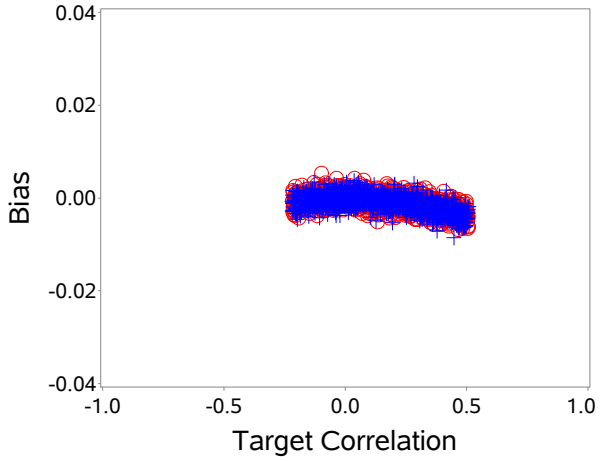
The proportion of simulations falling within the Fréchet bounds in the EP case started at 0.977, decreased slightly, then increased to a maximum of 0.991 at target correlation 0.374 (ρ^2 = 0.140), then decreased to 0.953. In the MS case, the proportion started at 0.976, decreased slightly, increased to 0.990 at target correlation 0.381 (ρ^2 = 0.145), then decreased to 0.954 at the upper bound.

The bias had a slight curve, with the greatest bias for both methods near the upper bound. For the EP case, this bias was -0.007, which occurred at target correlation 0.465 (ρ^2 = 0.216). In the MS case, the greatest bias was -0.009, occurring at target correlation 0.449 (ρ^2 = 0.210).

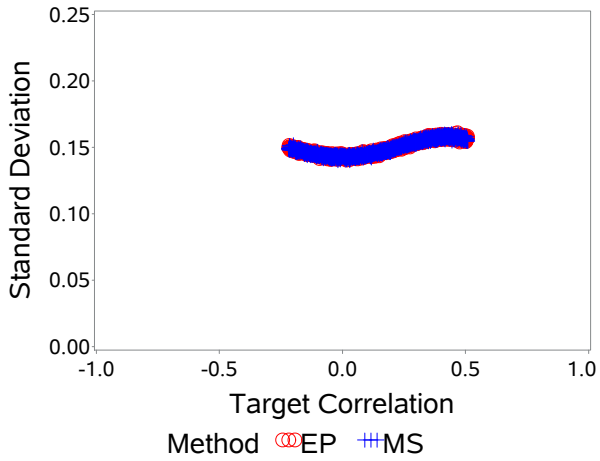
The standard deviation for the estimated correlation under the EP method started at 0.149, decreased slightly, then increased to a maximum of 0.161 at target correlation 0.466 (ρ^2 = 0.217), then decreased to 0.158 at the upper EP bound. In the MS case, the standard deviation started



(a) Proportion of simulations falling within the Fréchet bounds



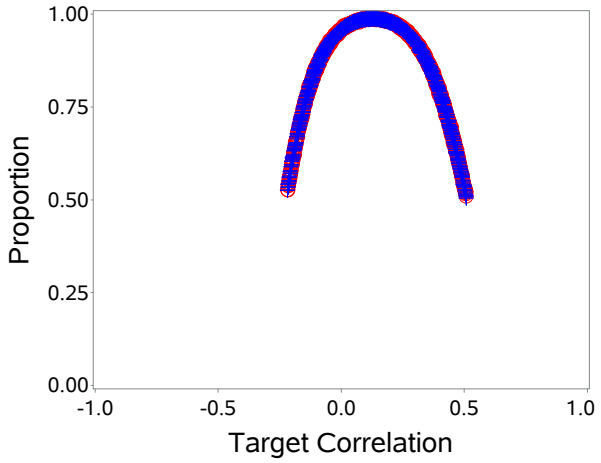
(b) Bias of the estimated correlation



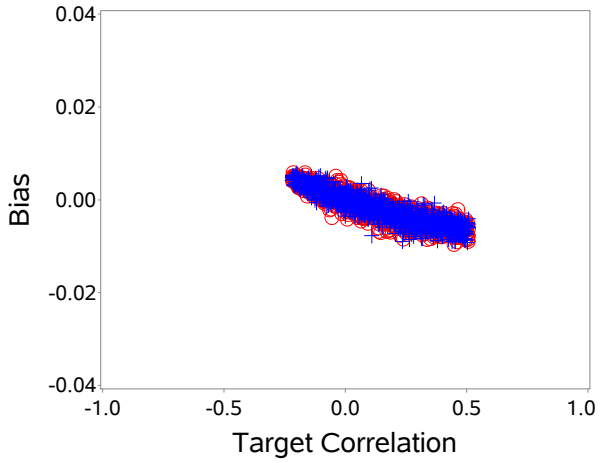
(c) Standard deviation of the estimated correlation

Fig. 2.29. Case (AR(1)): $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.1$. Figures for $\hat{\rho}_{13}$.

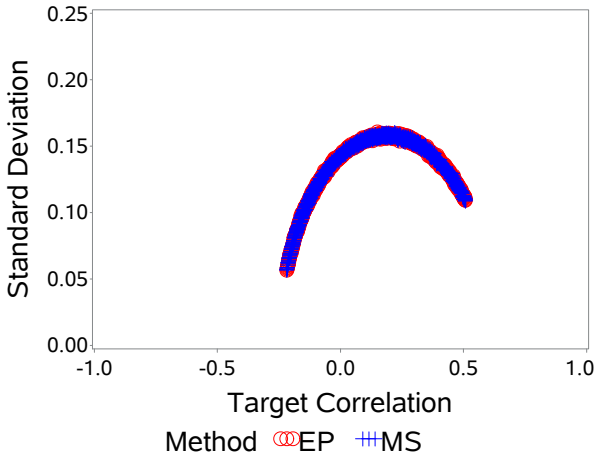
at 0.149 at the lower bound, decreased slightly, then increased to a maximum of 0.161 at target correlation 0.469 ($\rho^2 = 0.220$), then decreased back to 0.155 at the upper bound.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.30. Case (AR(1)): $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.1$. Figures for $\hat{\rho}_{23}$.

For $\hat{\rho}_{23}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.527 at the lower EP bound, increased to a maximum of 0.991 at target correlation 0.128, then dropped to 0.510 at the upper EP bound. For the MS case, the proportion started at 0.526 at the lower Fréchet bound, increased to a maximum of 0.991 at target correlation 0.148, then decreased back to 0.505 at the upper Fréchet bound.

The bias for both methods had a decreasing slope. The greatest bias for the EP method was -0.010, occurring at target correlation 0.447, and for the MS method, it was -0.009, occurring at target correlation 0.503.

The standard deviation for the estimated correlation under the EP method started at the EP lower bound at 0.057, increased to a maximum of 0.161 at target correlation 0.150, then decreased to 0.109 at the EP upper bound. In the MS case, the standard deviation started at 0.056 at the lower Fréchet bound, increased to 0.160 at target correlation 0.220, then decreased to 0.109 at the upper Fréchet bound.

Case (AR(1)): $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.3$

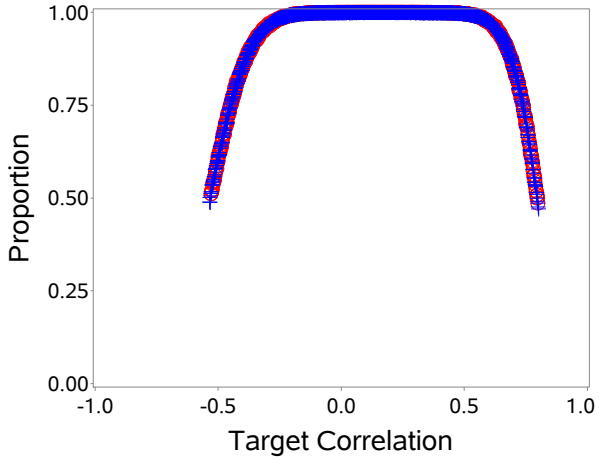
Representing a case where the marginal means steadily decrease, this case has Fréchet bounds [-0.535, 0.801]. The EP method again cannot estimate target correlations near the endpoints; the estimable range is [-0.529, 0.799].

For $\hat{\rho}_{12}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.511 at the lower EP bound, increased to a maximum of 1 throughout the middle of the range, then dropped to 0.486 at the upper EP bound. For the MS case, the proportion started at 0.489 at the lower Fréchet bound, increased to a maximum of 1 throughout the middle of the range, then decreased back to 0.471 at the upper Fréchet bound.

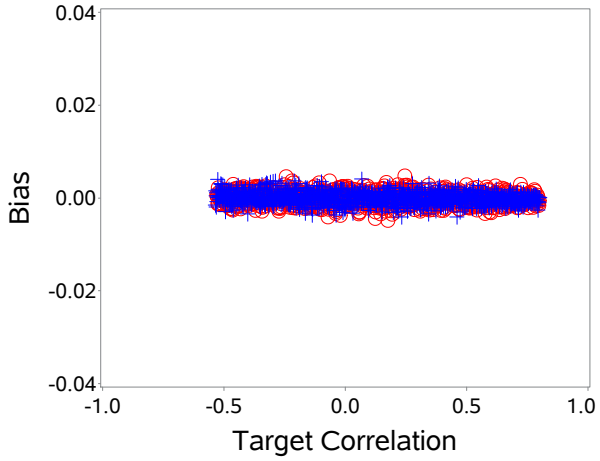
The bias for both methods stayed near zero. The greatest bias for the EP method was -0.005, occurring at target correlation 0.176, and for the MS method, it was -0.004, occurring at target correlation 0.233.

The standard deviation for the estimated correlation under the EP method started at the EP lower bound at 0.118, increased to a maximum of 0.145 at target correlation 0.029, then decreased to 0.076 at the EP upper bound. In the MS case, the standard deviation started at 0.118 at the lower Fréchet bound, increased to 0.145 at target correlation 0.106, then decreased to 0.074 at the upper Fréchet bound.

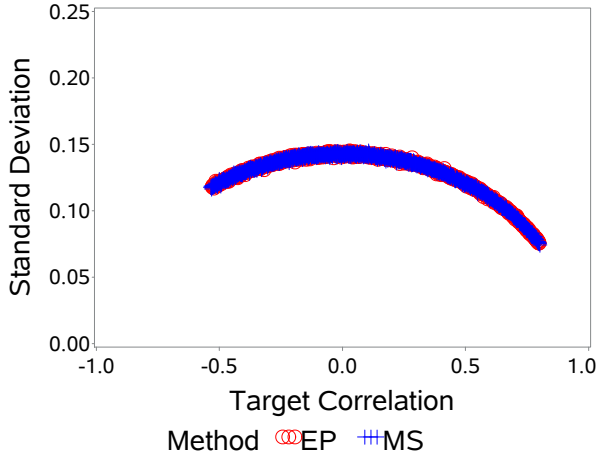
For $\hat{\rho}_{13}$, the quantity being estimated is ρ^2. For the sake of completion, estimates are presented



(a) Proportion of simulations falling within the Fréchet bounds

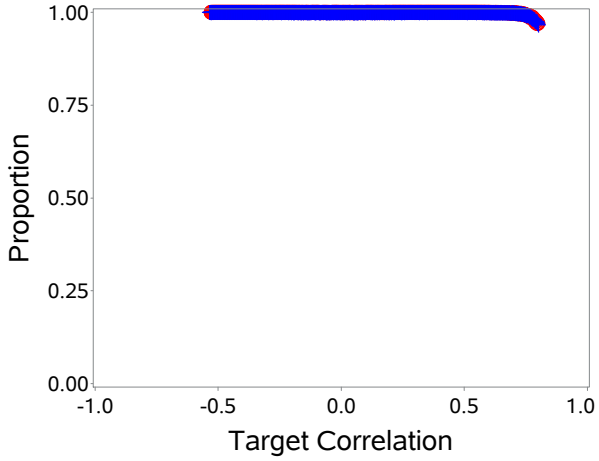


(b) Bias of the estimated correlation

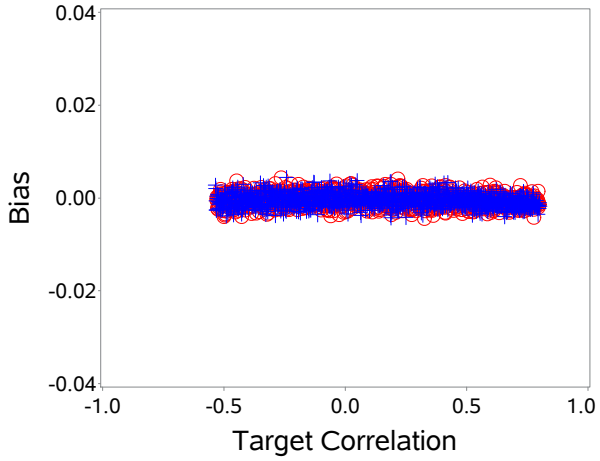


(c) Standard deviation of the estimated correlation

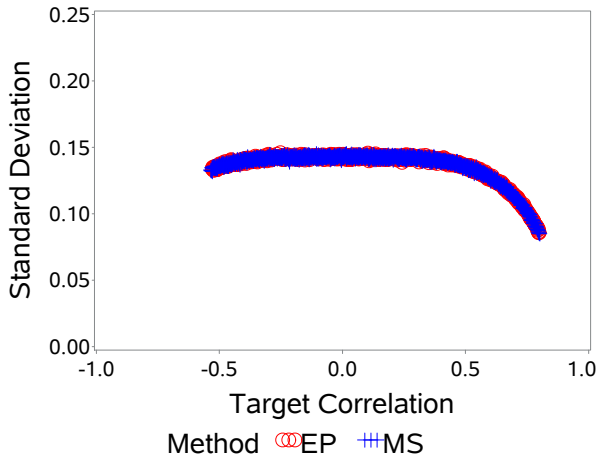
Fig. 2.31. Case (AR(1)): $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.3$. Figures for $\hat{\rho}_{12}$.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.32. Case (AR(1)): $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.3$. Figures for $\hat{\rho}_{13}$.

in reference to ρ. As such, the plots for the proportion and the standard deviation do not appear as partial parabolas.

The proportion of simulations falling within the Fréchet bounds in both cases started at 1, stayed there throughout the majority of the estimable range, then decreased to 0.971 at the EP upper bound in the EP case and to 0.967 at the upper Fréchet bound in the MS case.

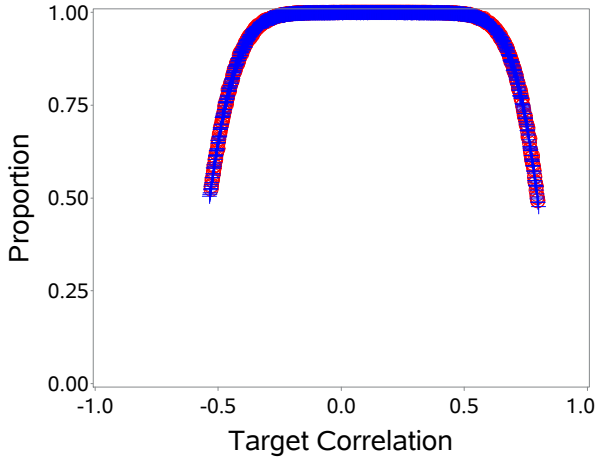
For the EP case, the maximum bias was -0.004, which occurred at target correlation 0.327 (ρ^2 = 0.107). In the MS case, the maximum was 0.004, occurring at target correlation -0.242 (ρ^2 = 0.059).

The standard deviation for the estimated correlation under the EP method started at 0.134, increased to a maximum of 0.146 at target correlation -0.252 (ρ^2 = 0.064), then decreased to 0.086 at the upper bound. In the MS case, the standard deviation started at 0.133 at the lower bound, increased to a maximum of 0.146 at target correlation 0.015 (ρ^2 = 0.0002), then decreased back to 0.085 at the upper bound.

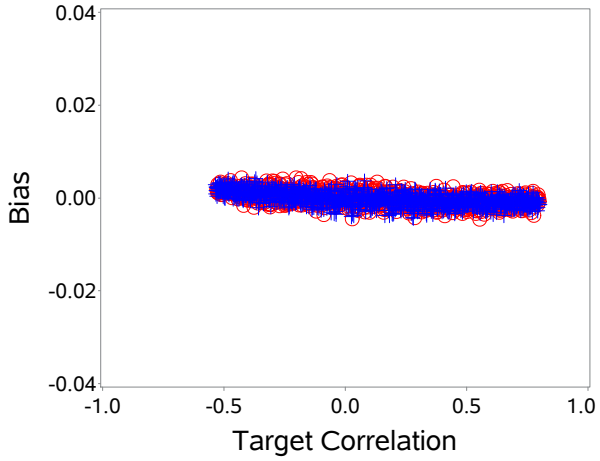
For $\hat{\rho}_{23}$, the proportion of simulations falling within the Fréchet bounds in the EP case started at 0.523 at the lower EP bound, increased to a maximum of 1 in the middle of the range, then dropped to 0.493 at the upper EP bound. For the MS case, the proportion started at 0.505 at the lower Fréchet bound, increased to a maximum of 1 and stayed there throughout the middle of the range, then decreased back to 0.477 at the upper Fréchet bound.

The bias for both methods stayed near zero. The greatest bias for the EP method was -0.005, occurring at target correlation 0.554, and for the MS method, it was -0.004, occurring at target correlation 0.280.

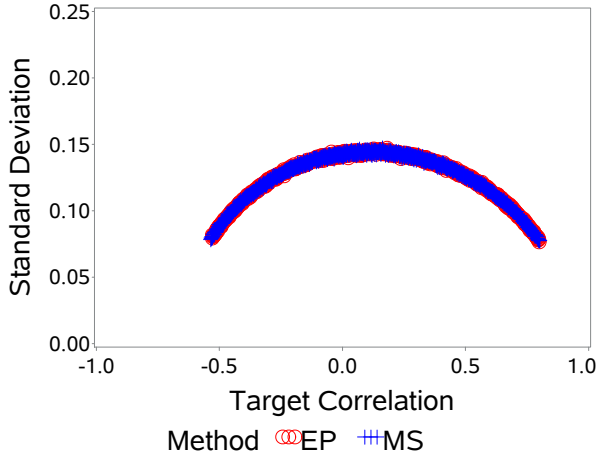
The standard deviation for the estimated correlation under the EP method started at the EP lower bound at 0.079, increased to a maximum of 0.147 at target correlation 0.181, then decreased to 0.077 at the EP upper bound. In the MS case, the standard deviation started at 0.078 at the lower Fréchet bound, increased to 0.162 at target correlation 0.147, then decreased to 0.078 at the upper Fréchet bound.



(a) Proportion of simulations falling within the Fréchet bounds



(b) Bias of the estimated correlation



(c) Standard deviation of the estimated correlation

Fig. 2.33. Case (AR(1)): $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.3$. Figures for $\hat{\rho}_{23}$.

Remarks Regarding Three-Measure Cases

In the three-measure cases, the MS method has a distinct advantage over the EP method. The EP method must have a positive definite correlation matrix (i.e. the matrix must be invertible) for the standard bivariate normal distribution to be used. This becomes a problem when the target correlation ρ_{ij} for a given combination of p_1, p_2, and p_3 does not allow for positive definiteness of $\Sigma = \left(\left(r_{ij}\right)\right)$. However, in general the results do not differ much between the two methods.

2.5 Simulated Two-Group Pre-/Post-Treatment Comparison

Using both the EP and MS methods, dependent data for two separate groups mimicking the case of a pre-/post-treatment comparison, as outlined in Table 2.2, where ρ is the correlation coefficient between pre- and post-treatment measurements in each group will be simulated. The null hypothesis of no difference in the change in the success rate from pre- to post-treatment between the two groups will be tested using the Generalized Estimating Equations (GEE) approach from Liang and Zeger [9] [11]. Comments on whether the estimated correlations fall within the Fréchet bounds and comparisons between the simulation methods will be made.

Another two-group comparison will be made with three test points (pre-treatment and two post-treatment tests) using an AR(1) correlation structure in order to explore the differences between the EP and MS methods in a three-variable case.

Table 2.2. Template for Two-Group Pre-/Post-Treatment Comparison

	Two-Test Case		Three-Test Case		
	p_1	p_2	p_1	p_2	p_3
Group 1	0.30	0.25	0.30	0.25	0.20
Group 2	0.30	0.10	0.30	0.15	0.10
n per group	100		250		

2.5.1 Methods

Using the EP and MS methods as described in Section 2.3, two sets of dependent binary data were simulated, one for Group 1 and one for Group 2 as described in Table 2.2, with $\rho = 0.4$ for each group. One thousand simulations were run for each method, starting with the same random seed (47). The working correlation–denoted $\hat{r}$–from GEE testing between the groups was recorded for comparison to the target correlation $\rho = 0.4$ and $\rho^2 = 0.16$ in the AR(1) case. The p-value (p) from the test of no difference in the change of success rate between the groups was also recorded. The percentage of working correlations which fall within the Fréchet bounds and the percentage of significant test results ($\alpha = 0.05$) were also calculated.

All calculations were completed using SAS 9.4 (The SAS Institute, Cary, NC). PROC IML was used for the simulating the correlated data, PROC CORR was used to estimate the correlations and means of the simulated variables, PROC GENMOD was used for testing the null hypothesis of no difference between the two groups at any time point under the GEE approach, PROC FREQ was used to determine percentages of correlations within the Fréchet bounds and results of significance testing, and PROC GPLOT was used for plotting results.

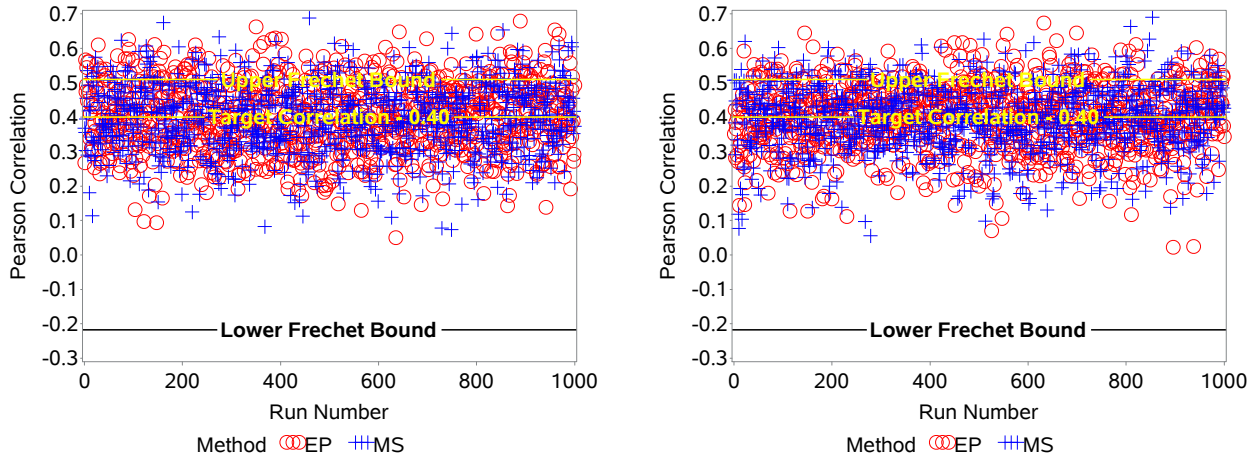
Note that in the three-test case, the group variable was modeled as a continuous variable in order to allow enough degrees of freedom to test the null hypothesis. This has no effect on the estimation or testing of the model as there are only two groups.

2.5.2 Two-Test Case Results

In cases where two sets of variables with differing marginal probabilities are also described as having the same correlation between the variables, the Fréchet bounds are those which are the most restrictive among the sets of marginal probabilities. The correlation cannot lie inside the Fréchet bounds for one set of p_js and not for the other. In this example, the Fréchet bounds were calculated to be [-0.218, 0.509].

After the variables were simulated using the EP and MS methods according to the specifications above, the Pearson correlation coefficient and empirical marginal probabilities were calcu-

lated from the raw data. The results were plotted according to run number. See Figure 2.34 and Figure 2.35.



(a) Group 1: Target = 0.40

(b) Group 2: Target = 0.40

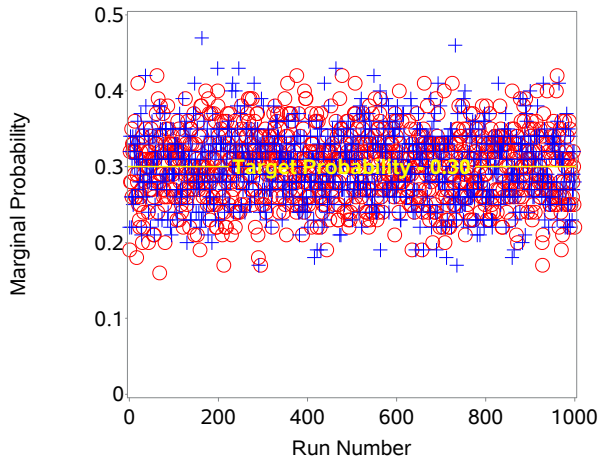
Fig. 2.34. Pearson Correlation between Pre- and Post-Treatment vs Run Number

The Pearson correlation for both methods and both groups varied between near zero to near 0.7. As the target correlation was 0.4, the spread seemed quite wide. However, with only 100 subjects per group, the calculated correlation was not expected to be perfectly simulated.

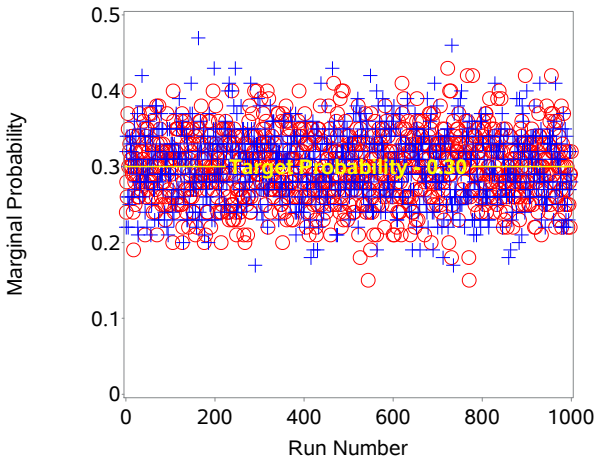
Table 2.3. Standard Deviations on Marginal Probabilities Across All Runs per Test per Group

		Standard Deviation	
Test	Group	EP	MS
Pre-treatment	1	0.0464	0.0455
	2	0.0452	0.0455
Post-treatment	1	0.0427	0.0421
	2	0.0302	0.0287

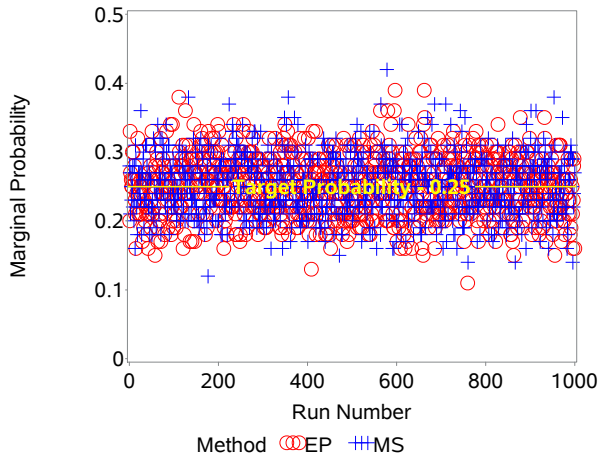
The marginal probability for each test was calculated by group, and the resulting estimates are shown in Figure 2.35. The standard deviations across the runs per test per method were calculated



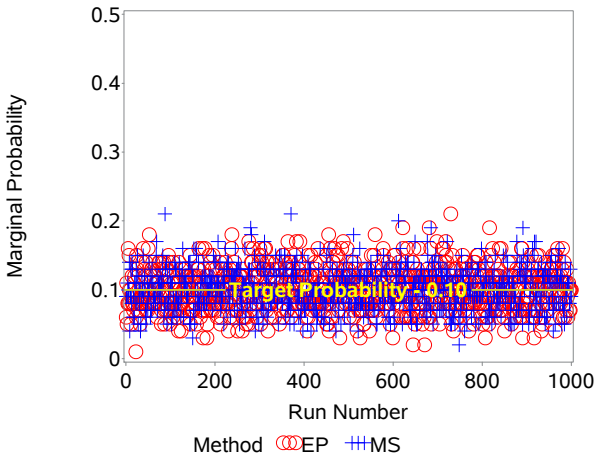
(a) Pre-Treatment in Group 1: Target = 0.30



(b) Pre-Treatment in Group 2: Target = 0.30



(c) Post-Treatment in Group 1: Target = 0.25



(d) Post-Treatment in Group 2: Target = 0.10

Fig. 2.35. Estimated Marginal Probability vs Run Number

to make a better comparison between the EP and MS results. This was estimated using PROC MEANS. As seen in Table 2.3, there was not much difference between the methods as far as the standard deviation of the estimated marginal probabilities.

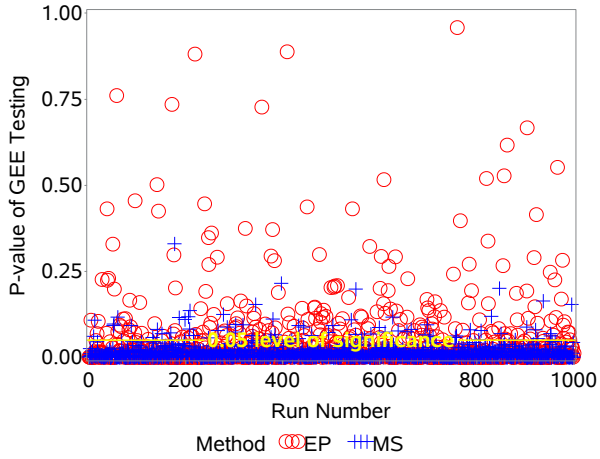


Fig. 2.36. P-value of GEE Testing vs Run Number: Target < 0.05

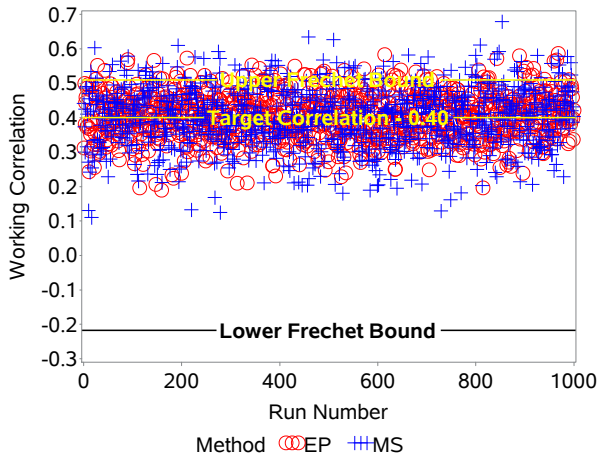


Fig. 2.37. GEE Working Correlation vs Run Number: Target = 0.40

After analyzing the simulated data using a GEE process in PROC GENMOD, it was found that the EP method fared worse in detecting the difference in the change of marginal probability between the groups. The EP method had significant test results in only 78.7% of the runs, whereas the MS method had significant results in 92.8% of runs, a difference of 14.1 percentage points.

The estimated correlation was within the Fréchet bounds for 93.8% of the runs for the EP method, and 90.5% of the runs for the MS method.

Remarks Regarding Two-Test Case Results

In the case of a pre-/post-treatment testing environment, it appears that the EP and MS methods are similar in regard to the estimation of the proper empirical correlation and marginal probabilities across a series of runs. However, the MS method is superior when it comes to detecting a difference between the groups under the GEE testing.

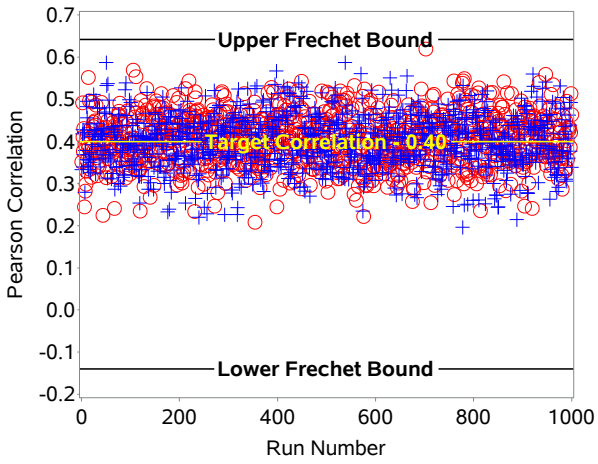
2.5.3 Three-Test Case Results

In the models where the correlation structure is of concern (i.e. cases with more than two variables), the Fréchet bounds change depending on said structure, as described by Chaganty and Joe (2006) [1]. According to their formulae for the AR(1) case, the most restrictive Fréchet bounds were [-0.140, 0.614].

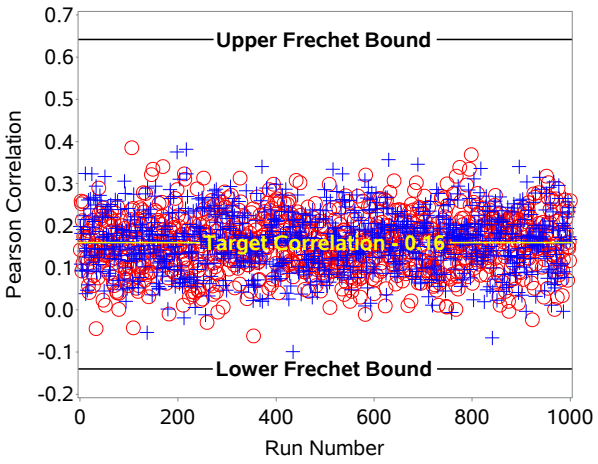
After the simulations were created by the EP and MS methods according to the specifications in Sections 2.5 and 2.5.1, the empirical values for the Pearson correlation and marginal probabilities were calculated. The results were plotted by run number. See Figures 2.38 and 2.39 for Pearson correlations.

The marginal probability for each test was calculated by group, and the resulting estimates are shown in Figure 2.40. The standard deviations across the runs per test per method were calculated to make a better comparison between the EP and MS results. This was estimated using PROC MEANS. As seen in Table 2.4, there was not much difference between the methods as far as the standard deviation of the estimated marginal probabilities or the appearance of the graphs.

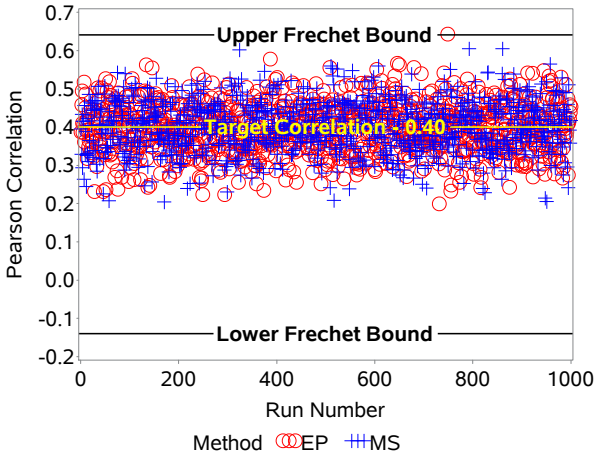
The GEE analysis procedure was much closer between the two methods in determining a difference between the two groups. The EP method had significant test results in 78.3% of the runs, and the MS method had significant test results in 80.8% of the runs. See Figure 2.41. Also, none of the simulated data resulted in a GEE working correlation that lay outside of the Fréchet



(a) ρ_{12}: Target = 0.40

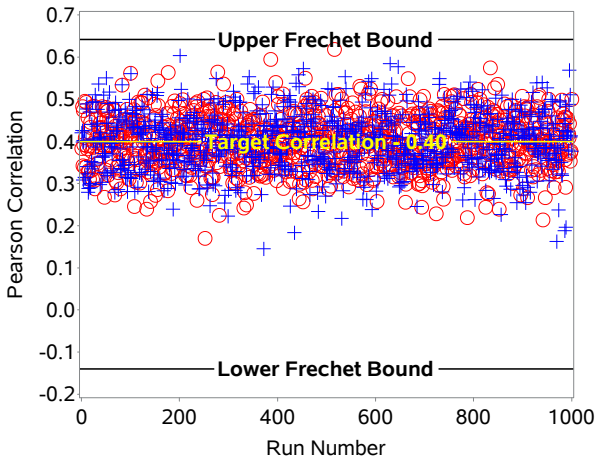


(b) ρ_{13}: Target = 0.16

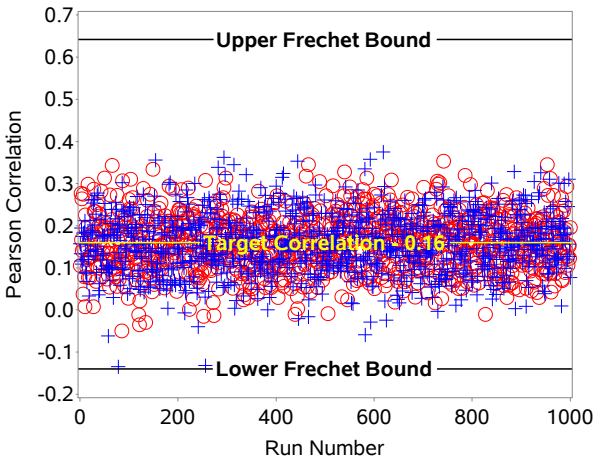


(c) ρ_{23}: Target = 0.40

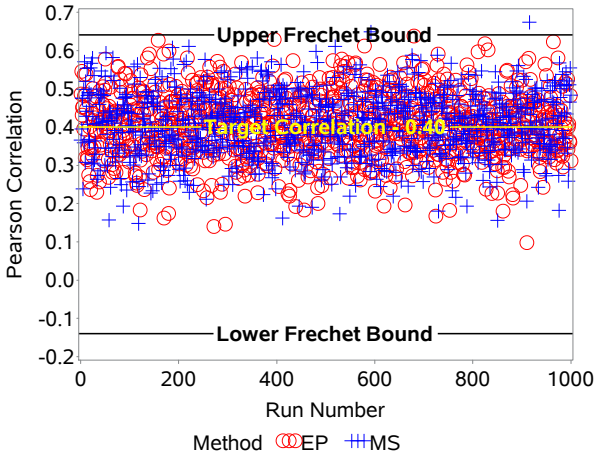
Fig. 2.38. Pearson Correlation between Pre- and Post-Treatment vs Run Number in Group 1



(a) ρ_{12}: Target = 0.40

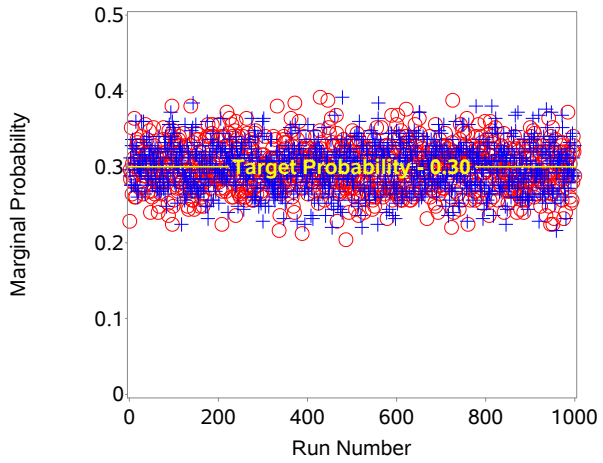


(b) ρ_{13}: Target = 0.16

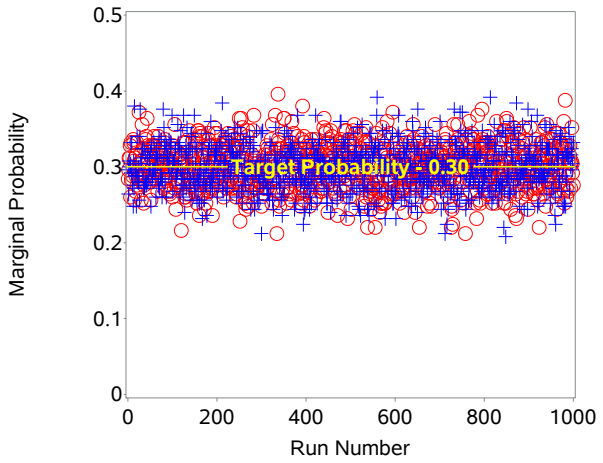


(c) ρ_{23}: Target = 0.40

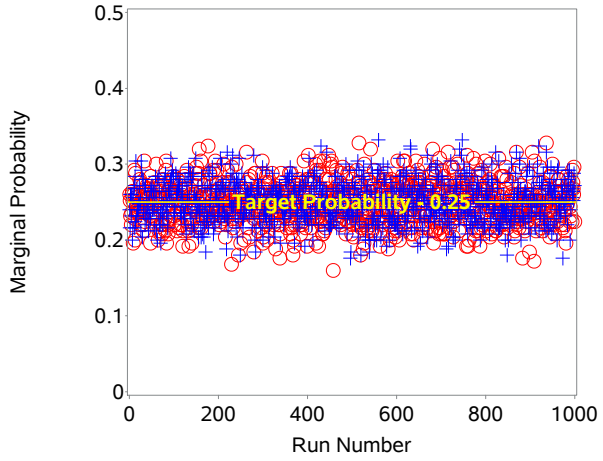
Fig. 2.39. Pearson Correlation between Pre- and Post-Treatment vs Run Number in Group 2



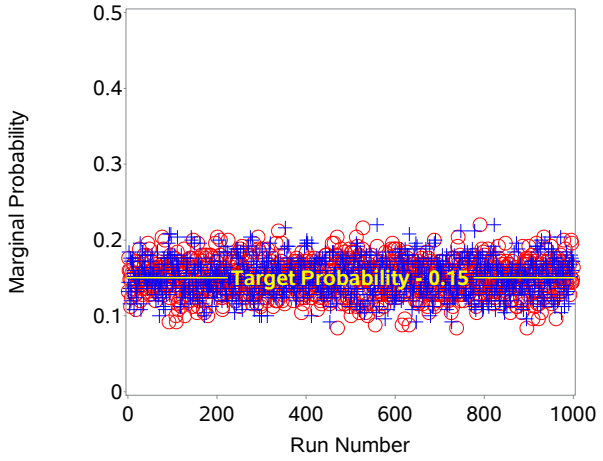
(a) Pre-Treatment in Group 1: Target = 0.30



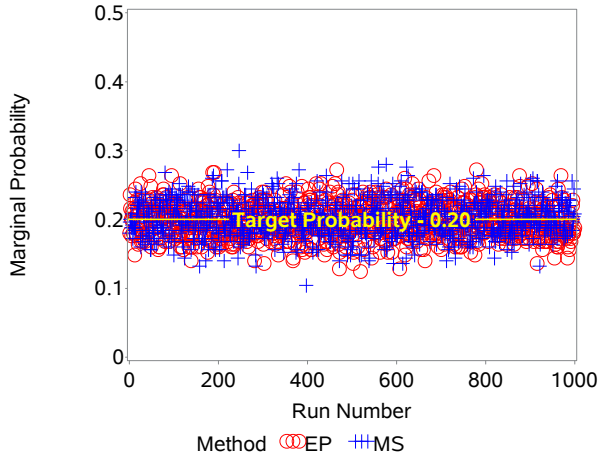
(b) Pre-Treatment in Group 2: Target = 0.30



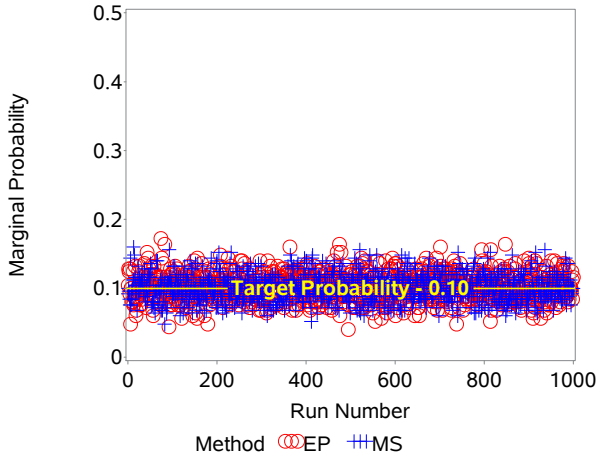
(c) First Post-Treatment in Group 1: Target = 0.25



(d) First Post-Treatment in Group 2: Target = 0.15



(e) Second Post-Treatment in Group 1: Target = 0.20



(f) Second Post-Treatment in Group 2: Target = 0.10

Fig. 2.40. Estimated Probability vs Run Number

Table 2.4. Standard Deviations on Marginal Probabilities Across All Runs per Test per Group

		Standard Deviation	
Test	Group	EP	MS
Pre-treatment	1	0.0300	0.0296
	2	0.0293	0.0293
Post-treatment 1	1	0.0274	0.0268
	2	0.0225	0.0222
Post-treatment 2	1	0.0250	0.0261
	2	0.0198	0.0193

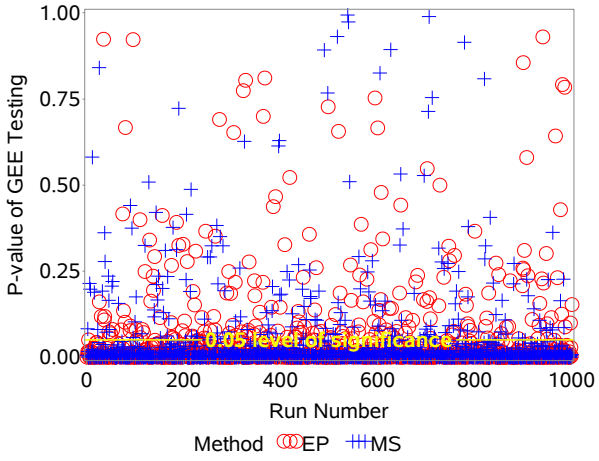


Fig. 2.41. P-value of GEE Testing vs Run Number: Target < 0.05

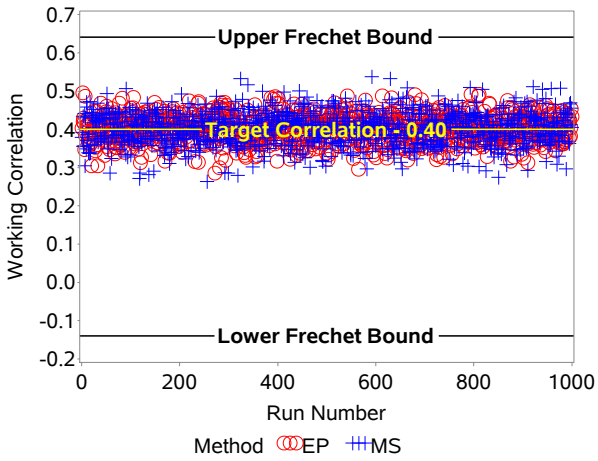


Fig. 2.42. GEE Working Correlation vs Run Number: Target = 0.40

bounds for either method, as seen in Figure 2.42.

Remarks Regarding Three-Test Case Results

In the case of a pre-/post-treatment testing environment with two post-treatment events, it appears that the EP and MS methods are similar in regard to the estimation of the proper empirical correlation and marginal probabilities across a series of runs. The MS method was better in this case at detecting the difference between the groups but only by a few percentage points.

CHAPTER 3

SIMULATING DEPENDENT BINARY VARIABLES USING THE ODDS RATIO

3.1 Introduction

The odds ratio is an appropriate measurement for many types of studies, making it of particular interest when examining dependence boundaries in relation to discrete data. This chapter will explore the ability of the multinomial sampling method to simulate data with desired odds ratios both in situations with a common odds ratio for all pairs of variables and in situations with potentially different odds ratios for all associations. The multinomial sampling method will be examined in detail according to certain measures of interest: the proportion of simulations needing adjustment to the odds ratio (to account for empty cells), the difference between the estimated mean odds ratio and the specified odds ratio (bias), and the standard deviation of the mean odds ratio.

The odds ratio is a measure of association between two variables, with an odds ratio greater than 1 indicating a positive association, and an odds ratio between zero and 1 indicating a negative association. For two binary variables Y_i and Y_j where the outcomes are 0 and 1, a count or "cell" n_{rs} (r, s = 0 or 1) is taken of each outcome, and a table is set up thus: The sample odds ratio

		Y_2	
		1	0
Y_1	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

is then calculated as $\psi_{ij} = n_{11}n_{00}/(n_{01}n_{10})$. This ψ is unbounded (i.e. $\psi \in [0, \infty]$) in the case of two binary variables, however, Fréchet bounds become a concern when more than two associated variables are being considered.

3.2 The Fréchet Bounds on the Odds Ratio

Recall the definitions of the Fréchet bounds from Chapter 1, especially Equation 1.1. The Fréchet bounds affect the odds ratio only in cases of three or more dependent variables. In a three-variable case, two of the odds ratios can vary in any manner, and the third is bounded. If a common odds ratio is assumed for all pairs of random variables Y_i and Y_j, then there may be a lower bound on the common odds ratio. Both types of Fréchet bounds are described in Chaganty and Joe (2006) [1]. From that paper, the following inequalities are obtained for the case of the common odds ratio and in the unstructured dependence case, respectively.

Theorem 2 from Chaganty and Joe [1] for the common odds ratio is as follows.

Let ψ_{ij} represent the odds ratio for a pair of binary random variables Y_i and Y_j. Let $p_{ij} = p_{ij}(\psi_{ij}) = C(p_i, p_j, \psi_{ij})$ where, for a fixed ψ, the function $C(u, v, \psi)$ is the Plackett copula, which for $0 \leq \psi < \infty$ is as follows:

$$C(p_i, p_j, \psi) = \frac{1 + (\psi - 1)(p_i + p_j) - \left[(1 + (\psi - 1)(p_i + p_j))^2 - 4\psi(\psi - 1)p_i p_j\right]^{1/2}}{2(\psi - 1)} \tag{3.1}$$

Consider three binary random variables Y_1, Y_2, and Y_3 with means p_1, p_2, and p_3. Assume a common odds ratio ψ for all pairs (Y_i, Y_j). A joint distribution for the three binary random variables exists if and only if

$$\psi_L(p_1, p_2, p_3) \leq \psi < \infty, \tag{3.2}$$

where $\psi_L(p_1, p_2, p_3) = 0$ if $p_1 + p_2 + p_3 \leq 1$ or if $p_1 + p_2 + p_3 \geq 2$. For $1 < p_1 + p_2 + p_3 < 2$, $\psi_L(p_1, p_2, p_3)$ is the positive root of the equation $p_{12}(\psi) + p_{13}(\psi) + p_{23}(\psi) - p_1 - p_2 - p_3 + 1 = 0$. Recall $p_{ij}(\psi) = C(p_i, p_j, \psi)$ from above. All other quantities are as defined in Chapter 2 Section 2.2.

The bounds for odds ratios in a three-variable unstructured case are described by Chaganty and Joe [1] as follows. Let the odds ratios for the three bivariate combinations of the variables are ψ_{12}, ψ_{13}, and ψ_{23}. Let ψ_{12} and ψ_{23} be unconstrained, taking values in the region $[0, \infty)$ independently. For ψ_{13} to be compatible with the marginal probabilities p_1, p_2, and p_3, the necessary and

sufficient range for ψ_{13} is given by

$$\psi_{13L} = \frac{p_{13L}(1-p_1-p_3+p_{13L})}{(p_1-p_{13L})(p_3-p_{13L})} \leq \psi_{13} \leq \frac{p_{13U}(1-p_1-p_3+p_{13U})}{(p_1-p_{13U})(p_3-p_{13U})} = \psi_{13U} \tag{3.3}$$

where, according to equation (5) from Chaganty and Joe [1],

$$p_{13L} = \max\{0,\, p_{12}+p_{23}-p_2,\, p_1+p_2+p_3-p_{12}-p_{23}-1,\, p_1+p_3-1\} \tag{3.4}$$

$$p_{13U} = \min\{p_1,\, p_3,\, p_1+p_{23}-p_{12},\, p_3+p_{12}-p_{23}\}. \tag{3.5}$$

3.3 The Common Odds Ratio Case

Utilizing Theorem 2 from Chaganty and Joe [1] (Equation 3.2), the lower bound for a series of three-variable marginal probability combinations was found and simulations were run to determine the accuracy of the multinomial sampling method in creating datasets with a target odds ratio. The Emrich and Piedmonte method (see Section 2.3) cannot be used to create datasets using the odds ratio since the method relies on the correlation to create a multivariate normal distribution, and there is no direct association between the correlation and the odds ratio. Thus, there cannot be a comparison between the two methods.

3.3.1 Methods

In order to make use of Equation 3.2, the Plackett copula $p_{ij}(\psi)$ must be calculated for each pair of marginal probabilities. Using the positive root of the equation $p_{12}(\psi)+p_{13}(\psi)+p_{23}(\psi)-p_1-p_2-p_3+1=0$ for ψ_L, a series of 100 target odds ratios was created, from the lower bound up to a maximum of 100 for four separate cases. The upper bound of 100 was chosen as an approximation for infinity. These target odds ratios were chosen by taking the log (base 10) of the endpoints–using 0.001 as an approximation for 0–and subdividing the resulting interval into 100 points. These points were then exponentiated back to the original scale and used as the target odds ratios.

The joint probability was also calculated using the Plackett copula, (recall $p_{ij} = C(p_i, p_j, \psi_{ij})$)

If the result for p_{13} did not satisfy $p_{13L} \leq p_{13} \leq p_{13U}$ according to Equations 3.4 and 3.5, the joint probability for all three variables p_{123} could not be calculated and the run was regarded as missing.

The number of observations simulated was 100, with 10,000 iterations used to estimate the properties of interest: proportion of simulations where all odds ratios fell within the Fréchet bounds, proportion of simulations with odds ratio adjustment (i.e., proportion of odds ratios with a zero cell), bias, and efficiency as measured by standard deviation. Results with a zero cell were adjusted using a commonly employed technique, adding 0.5 to each cell in order to make the odds ratio tractable. All results were calculated using SAS 9.4 (The SAS Institute, Cary, NC) in PROC IML. Figures were produced using PROC GPLOT.

3.3.2 Results

Four cases were chosen and the lower Fréchet bound was calculated, as shown in Table 3.1. The common odds ratio case assumes $\psi_{12} = \psi_{13} = \psi_{23}$, so results will be presented for each ψ_{ij}, in order to allow comparison between the three.

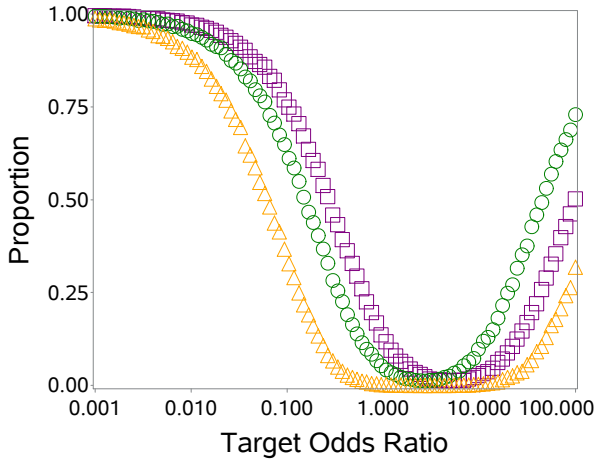
Table 3.1. Cases, Lower Bounds, and Upper Bounds for Common Odds Ratio ψ

p_1, p_2, p_3	Lower Bound	Upper Bound
0.1, 0.2, 0.3	0	∞
0.3, 0.4, 0.5	0.155	∞
0.5, 0.6, 0.7	0.155	∞
0.7, 0.8, 0.9	0	∞

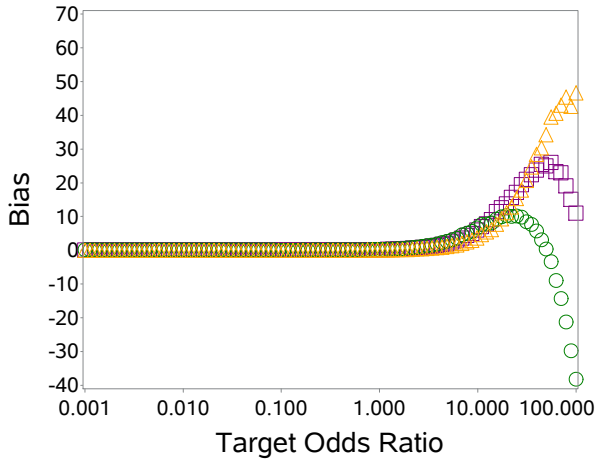
Case: $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$

Since zero was the lower bound for this case, there were no calculated odds ratios outside of the Fréchet bounds.

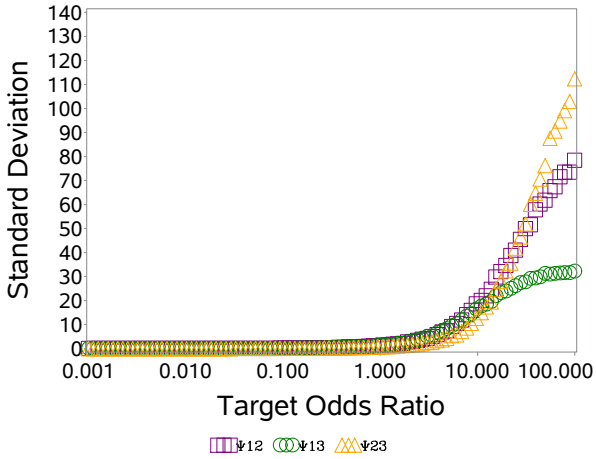
The proportion of simulations with a zero cell in need of adjusting the cells in order to calcu-



(a) Proportion of Simulations with Adjusted Odds Ratio



(b) Bias of Mean Odds Ratio Estimate



(c) Standard Deviation of Mean Odds Ratio Estimate

Fig. 3.1. Case: $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$ Plots for Measures of Interest

late the odds ratio varied between the three estimates of ψ. Since the target odds ratio had a lower bound of 0, it was not unexpected that all or nearly all simulations required an adjustment to the cells at that point. For all three ψ_{ij}s, the proportion needing adjustment followed a general trend, starting at 1, then dropping sharply to 0, and rising again as the target odds ratio increased past 1.

As the odds ratio increased, so did the bias up until near 60 for ψ_{12} and about 20 for ψ_{13}, after which the bias decreases. For ψ_{13}, the bias becomes negative near 50, decreasing to -38.2 at target odds ratio 100. The bias was never negative for ψ_{12} and ψ_{23}, however, this is probably due to where the upper bound was chosen. See Figure 3.1b.

The standard deviation also generally increased as the target odds ratio increased, reaching about 78.5 for ψ_{12}, 32.2 for ψ_{13}, and a little over 112.4 in the case of ψ_{23}. The standard deviation increased more dramatically after the target odds ratio reached about 3. See Figure 3.1c.

Generating data based on the odds ratio is inherently unstable as the odds ratio is not necessarily a direct association between two variables. Only the marginal probabilities need to remain constant, and so the joint probabilities are less restricted than in the case of generating data using the correlation.

Case: $p_1 = 0.3$, $p_2 = 0.4$, $p_3 = 0.5$

The lower bound ψ_L for this case was 0.155. Right at that boundary, the proportion of simulations with the odds ratio inside the bound was only 0.025. This increased to 0.995 as the target odds ratio increased to 0.696, reaching 1 starting at 1.254. See Figure 3.2.

The proportion of simulations in need of adjustment to the odds ratio did not vary much between the three estimates of ψ. See Figure 3.3a. For all three ψ_{ij}s, the proportion needing adjustment followed a general trend, starting near or at zero at the bound, then dropping to 0, and rising slowly as the target odds ratio increased past about 4.2. The increase becomes steeper as the odds ratio increases past about 20. Note that ψ_{13} has a noticeably steeper trend than the other estimates after this point.

As in the previous case, as the odds ratio increased, so did the bias. From Figure 3.3b, note

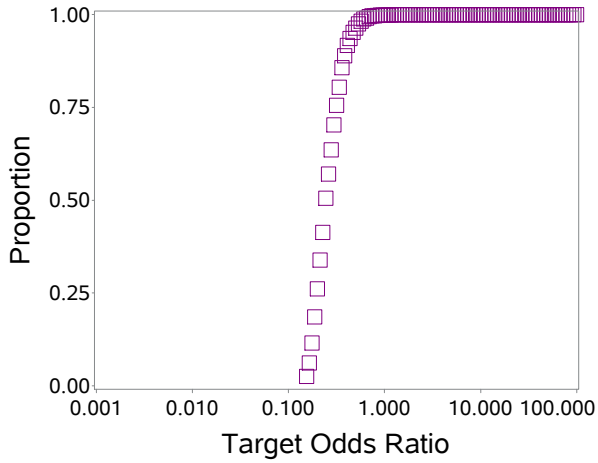


Fig. 3.2. Case: $p_1 = 0.3$, $p_2 = 0.4$, $p_3 = 0.5$ Proportion of Simulations with Odds Ratio Within the Fréchet Bounds

that the bias appears to plateau and decrease after the target odds ratio increased above about 35 for ψ_{13}, decreasing to 5.055 at 100. There is little difference between the trends for ψ_{12} and ψ_{23} and both increase to approximately 60 at target odds ratio 100.

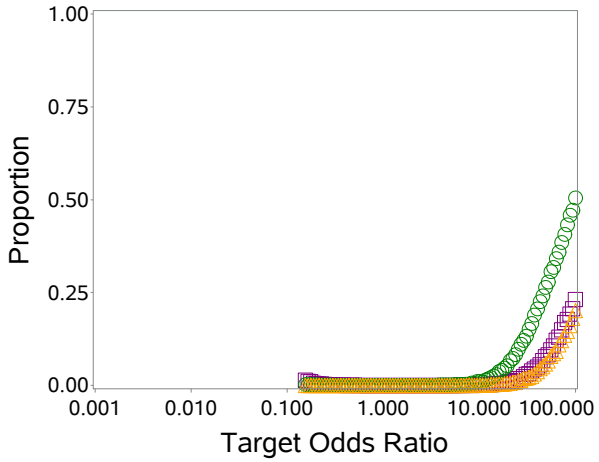
The standard deviation also generally increased as the target odds ratio increased, reaching close to 130 for ψ_{12} and ψ_{23}. For ψ_{13}, this only increased to 60. See Figure 3.3c.

Case: $p_1 = 0.5$, $p_2 = 0.6$, $p_3 = 0.7$

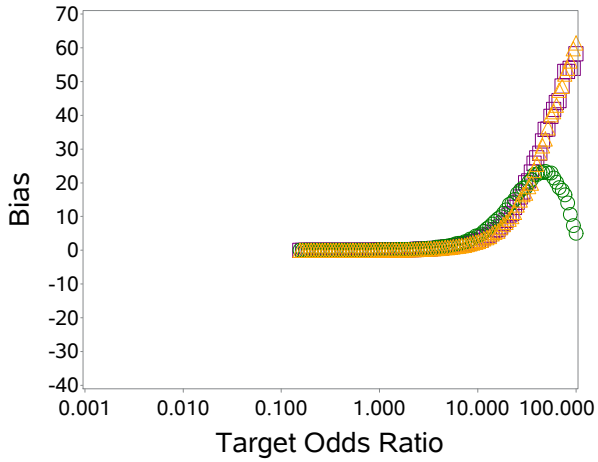
The lower bound ψ_L for this case was 0.155. Right at that boundary, the proportion of simulations with the odds ratio within the bounds was only 0.024. This increased drastically to 0.999 at target odds ratio 0.848. See Figure 3.4.

The proportion of simulations in need of adjustment to the odds ratio did not vary much between the three estimates of ψ. For all three ψ_{ij}s, the proportion needing adjustment followed a general trend, starting near or at zero at the bound, then dropping to 0, and rising slowly as the target odds ratio increased past about 4.2. The increase becomes steeper as the odds ratio increases past about 20. Note that ψ_{13} has a noticeably steeper trend after this point. See Figure 3.5a.

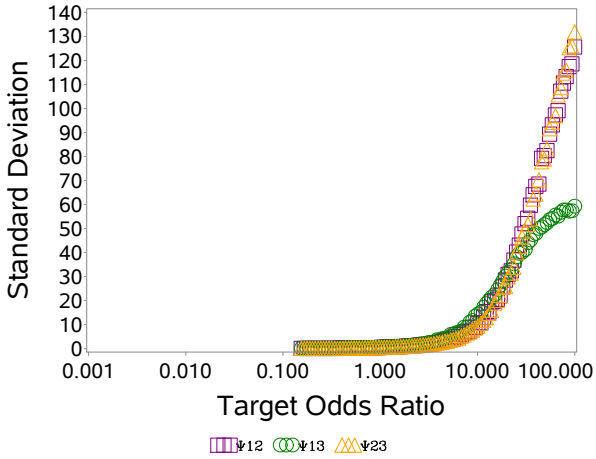
As in the previous case, as the odds ratio increased, so did the bias, but only in the positive



(a) Proportion of Simulations with Adjusted Odds Ratio



(b) Bias of Mean Odds Ratio Estimate



(c) Standard Deviation of Mean Odds Ratio Estimate

Fig. 3.3. Case: $p_1 = 0.3$, $p_2 = 0.4$, $p_3 = 0.5$ Plots for Measures of Interest

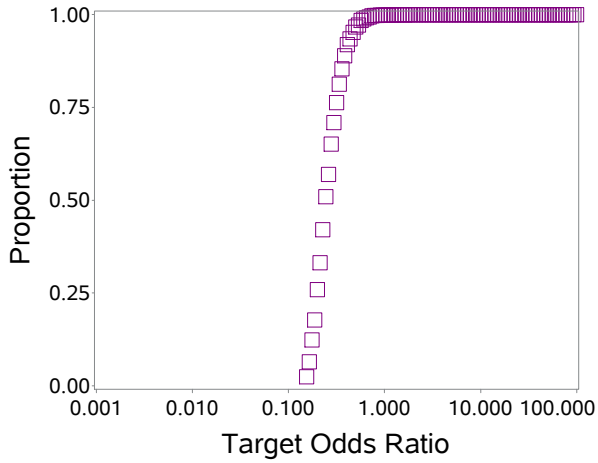


Fig. 3.4. Case: $p_1 = 0.5$, $p_2 = 0.6$, $p_3 = 0.7$ Proportion of Simulations with Odds Ratio Within the Fréchet Bounds

direction. From Figure 3.5b, note that the bias appears to decrease after the target odds ratio increased above about 35 for ψ_{13}, down to 4.32 at target odds ratio 100. There is little difference between the trends for ψ_{12} and ψ_{23}, both reaching about 60 at target odds ratio 100.

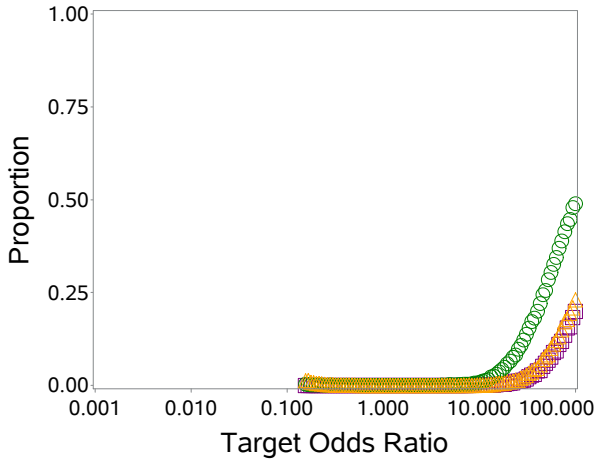
The standard deviation also generally increased as the target odds ratio increased, reaching around 130 for ψ_{12} and ψ_{23} and around 60 for ψ_{13}. See Figure 3.5c.

Case: $p_1 = 0.7$, $p_2 = 0.8$, $p_3 = 0.9$

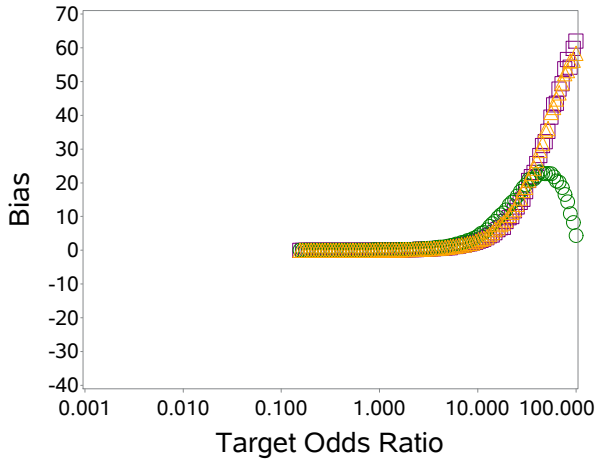
The lower bound ψ_L for this case was 0, so there were no estimates outside of the bounds.

The proportion of simulations of adjusting the odds ratio varied between the three estimates of ψ. Since the target odds ratio had a lower bound of 0, it was not unexpected that all or nearly all simulations required an adjustment to the cells at that point. For all three ψ_{ij}s, the proportion needing adjustment followed a general trend, starting at 1, then dropping sharply to 0, and rising again as the target odds ratio increased past 1. See Figure 3.6a.

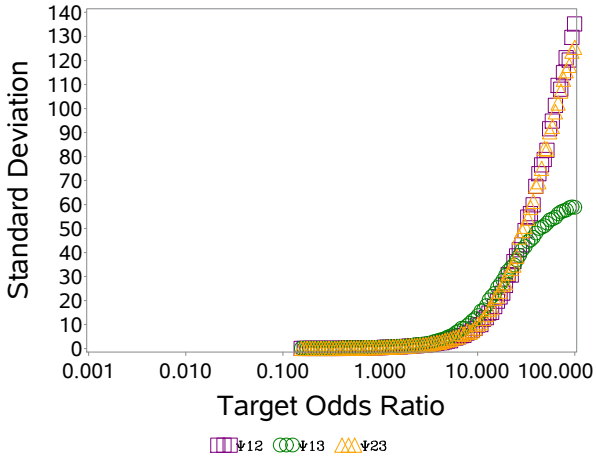
From Figure 3.6b, note that ψ_{12} increases until 100, but it appears that it begins to have a downward turn, even so reaching a maximum of 46.04. The bias began decreasing after the target odds ratio increased above about 25 for ψ_{13}, having reached a maximum of about 11. At target



(a) Proportion of Simulations with Adjusted Odds Ratio

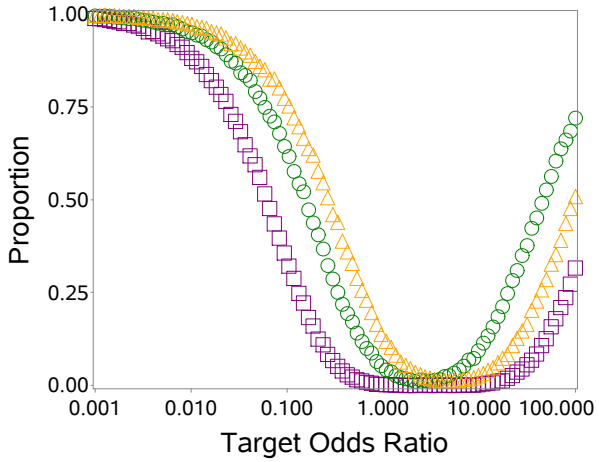


(b) Bias of Mean Odds Ratio Estimate

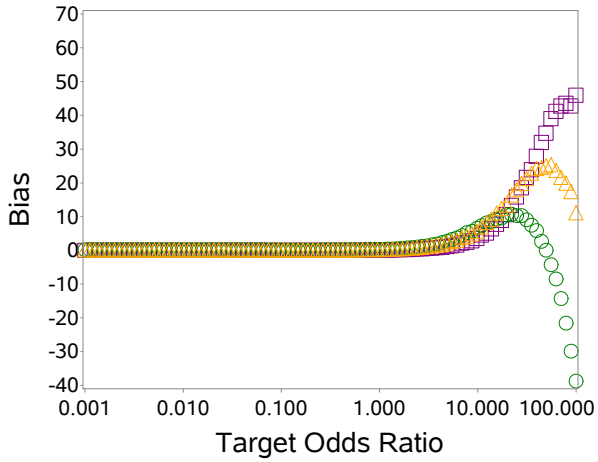


(c) Standard Deviation of Mean Odds Ratio Estimate

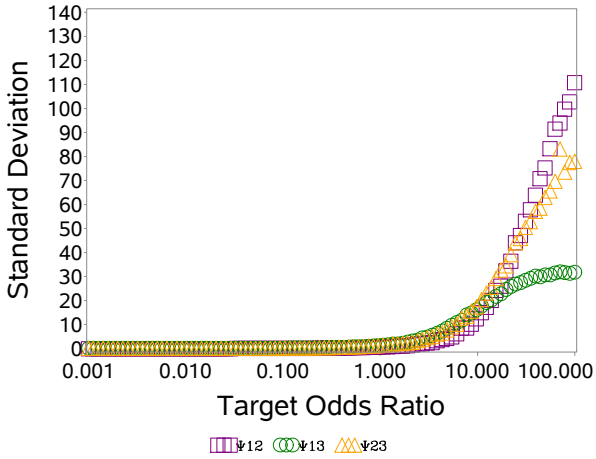
Fig. 3.5. Case: $p_1 = 0.5$, $p_2 = 0.6$, $p_3 = 0.7$ Plots for Measures of Interest



(a) Proportion of Simulations with Adjusted Odds Ratio



(b) Bias of Mean Odds Ratio Estimate



(c) Standard Deviation of Mean Odds Ratio Estimate

Fig. 3.6. Case: $p_1 = 0.7$, $p_2 = 0.8$, $p_3 = 0.9$ Plots for Measures of Interest

odds ratio 100, this reached a minimum of -38.8. The bias for ψ_{23} reached a maximum of 25.5 at target odds ratio 55.91, dropping to 11.14 at 100.

The standard deviation also generally increased as the target odds ratio increased, reaching about 110 for ψ_{12}, 32 for ψ_{13}, and 78 for ψ_{23}. See Figure 3.6c.

Remarks Regarding Common Odds Ratio Case

It is interesting to note that in the pairs of cases with marginal probabilities symmetric about 0.5 (e.g. the case where p_1=0.1, p_2=0.2, and p_3=0.3 and the case where p_1=0.7, p_2=0.8, and p_3=0.9), the measures for ψ_{12} behaved the same as for ψ_{23} and vice versa, while ψ_{13} remained the same in both cases.

3.4 The Unstructured Odds Ratio Case

Using the unstructured portion of Section 5 of Chaganty and Joe [1], the upper and lower Fréchet bounds on ψ_{13} were calculated for three marginal probability combinations with specified ψ_{12} and ψ_{23}. Simulations were performed to observe the accuracy of the multinomial sampling method in creating datasets with a target odds ratio. The Emrich and Piedmonte method (see Section 2.3) cannot be used to create datasets using the odds ratio since the method relies on the correlation to create a multivariate normal distribution, and there is no direct association between the correlation and the odds ratio. Thus, again, there cannot be a comparison between the two methods.

3.4.1 Methods

The Plackett copula was used to calculate the joint probability function from the specified marginal probabilities and associated odds ratios ψ_{12} and ψ_{23}. The joint probability function was used along with Equations 3.3-3.5 to determine the bounds on ψ_{13}. If the upper bound was not calculable according to Equation 3.3 (i.e. if $p_{13U} = p_1$ or p_3), the upper bound was defined as infinity. After the bounds were determined, a series of target odds ratios was created, from the

lower bound to the upper bound. These target odds ratios were chosen by taking the log (base 10) of the endpoints–using 0.001 as an approximation for 0 and 100 as an approximation for infinity– and subdividing the resulting interval into 100 points. These points were then exponentiated back to the original scale and used as the target odds ratios.

The number of observations simulated was 100, with 10,000 iterations used to estimate the properties of interest: proportion of simulations where the estimated odds ratio fell within the Fréchet bounds, the proportion of simulations with odds ratio adjustment, the mean bias, and the efficiency (standard deviation). Results needing an adjustment to the odds ratio were those with a zero cell. The odds ratio was subsequently adjusted using a common technique: adding 0.5 to each cell before calculating the ratio. All results were calculated using SAS 9.4 (The SAS Institute, Cary, NC) in PROC IML. Figures were produced using PROC GPLOT.

3.4.2 Results

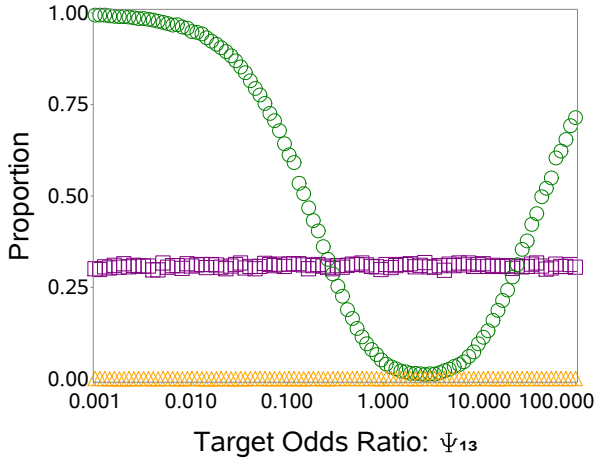
Three cases were chosen as representative of the possibilities for the bounds: zero to infinity, the lower bound greater than zero and a finite upper bound, and zero to a finite upper bound. The three marginal probabilities, ψ_{12}, and ψ_{23} were fixed. The Fréchet bounds on ψ_{13} were calculated, as shown in Table 3.2. For these cases, ψ_{13} was allowed to vary across the entire range within the bounds, and the measures of interest were calculated for each ψ_{ij}.

Table 3.2. Cases, Lower Bounds, and Upper Bounds for Unstructured Odds Ratio

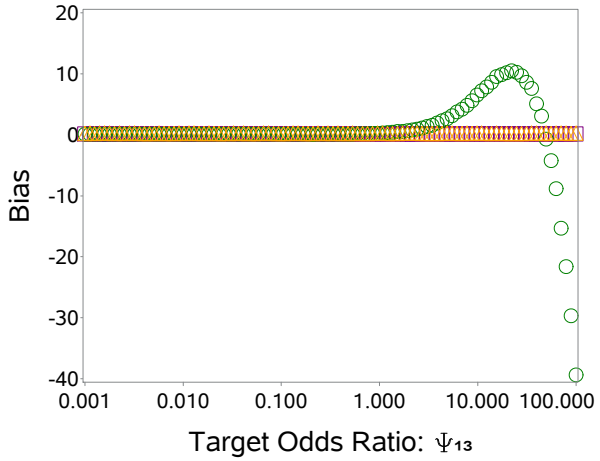
			ψ_{13}	ψ_{13}
p_1, p_2, p_3	ψ_{12}	ψ_{23}	Lower Bound	Upper Bound
0.1, 0.2, 0.3	0.50	1.75	0	∞
0.5, 0.4, 0.55	8.00	1.50	0.13	19.77
0.6, 0.5, 0.6	0.50	1.75	0	30.86

Case: $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$; $\psi_{12} = 0.50$, $\psi_{23} = 1.75$

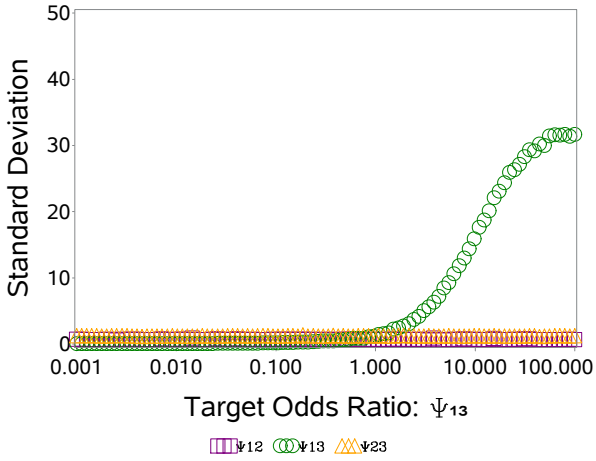
Since the Fréchet bounds in this case are $[0, \infty)$, the whole range for odds ratios in general, there were no estimated odds ratios outside of the bounds.



(a) Proportion of Simulations with Adjusted Odds Ratio



(b) Bias of Mean Odds Ratio Estimate



(c) Standard Deviation of Mean Odds Ratio Estimate

Fig. 3.7. Case: $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$; $\psi_{12} = 0.50$, $\psi_{23} = 1.75$ Plots for Measures of Interest

The proportion of simulations needing an adjustment to the odds ratio due to a zero cell varied according to which odds ratio was being measured. For ψ_{12}, which was relatively close to zero at 0.50, needed adjustment about every 3 out of 10 simulations, and ψ_{23} hardly ever needed

adjustment. For ψ_{13}, the proportion with a zero cell started off at 1, since the target odds ratio was near zero, decreased to a minimum of 0.01 at target odds ratio 2.53, then rose again steadily to 0.71 at 100.

The bias was positive but low for ψ_{12} (≈ 0.18) and ψ_{23} (≈ 0.27). For ψ_{13}, the bias stayed steady around 0.1 until target odds ratio 0.37, when it began to increase to 10.5 at 22.05, then decreasing to -39.4 at 100.

The standard deviation stayed steady for ψ_{12} and ψ_{23}, as might be expected from the previous results. For ψ_{12}, the standard deviation stayed near 0.72, and for ψ_{23} it was near 1.23. For ψ_{13}, the deviation increased from near 0 starting around target odds ratio 1, up until 31.67 at 100.

Case: $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.55$; $\psi_{12} = 8.00$, $\psi_{23} = 1.50$

The Fréchet bounds in this case are $[0.125, 19.772]$. None of the measures of interest were able to be calculated at the lower bound due to the estimate for p_{13} per the Placket copula (Equation 3.1) lying outside of the bounds given by Equations 3.4 and 3.5.

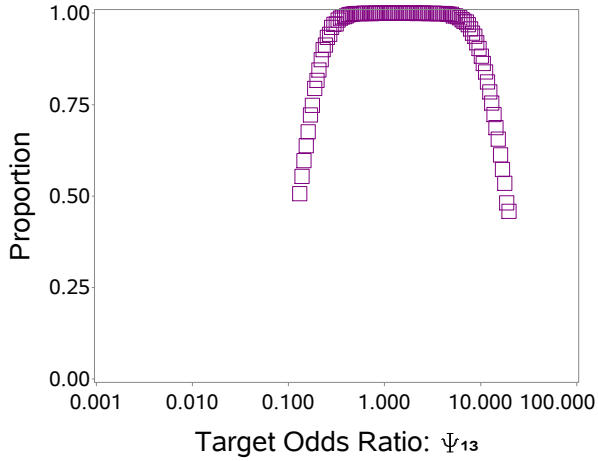
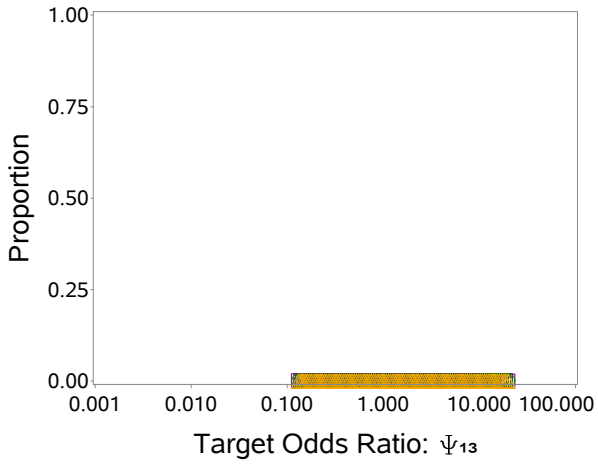


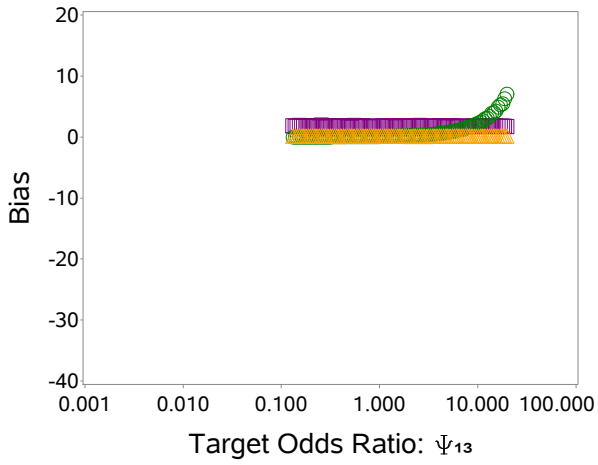
Fig. 3.8. Case: $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.55$; $\psi_{12} = 8.00$, $\psi_{23} = 1.50$ Proportion of Simulations with Odds Ratio Within the Fréchet Bounds

The proportion of simulations with the estimated odds ratio falling within the bounds started at 0.507 at target odds ratio 0.131, increased to 1 by 0.789, started decreasing at 2.193, and ended

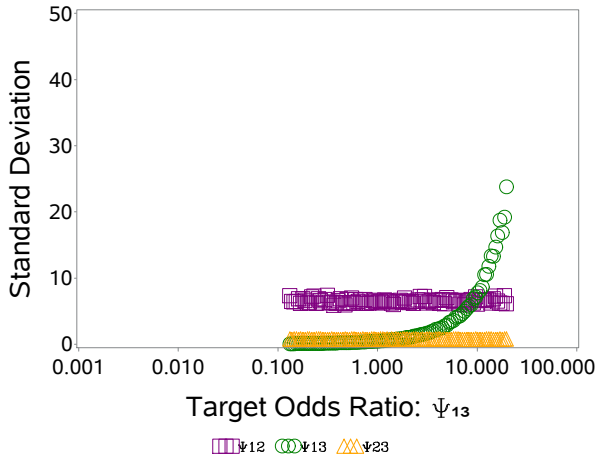
at 0.459 at the upper bound. See Figure 3.8.



(a) Proportion of Simulations with Adjusted Odds Ratio



(b) Bias of Mean Odds Ratio Estimate



(c) Standard Deviation of Mean Odds Ratio Estimate

Fig. 3.9. Case: $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.55$; $\psi_{12} = 8.00$, $\psi_{23} = 1.50$ Plots for Measures of Interest

The proportion of simulations needing an adjustment to the odds ratio due to a zero cell was negligibly different from zero across all target odds ratios.

The bias was positive but fairly low for ψ_{12} (≈ 1.8) and ψ_{23} (≈ 0.16). For ψ_{13}, the bias increased fairly steadily from near 0 to 7.02 at the upper bound.

For ψ_{12}, the standard deviation stayed near 6.7, and for ψ_{23} it was near 0.76. For ψ_{13}, the deviation increased from 0.07 starting at target odds ratio 0.131, up until 23.79 at the upper Fréchet bound.

See Figure 3.9 for plots.

Case: $p_1 = 0.6$, $p_2 = 0.5$, $p_3 = 0.6$; $\psi_{12} = 0.50$, $\psi_{23} = 1.75$

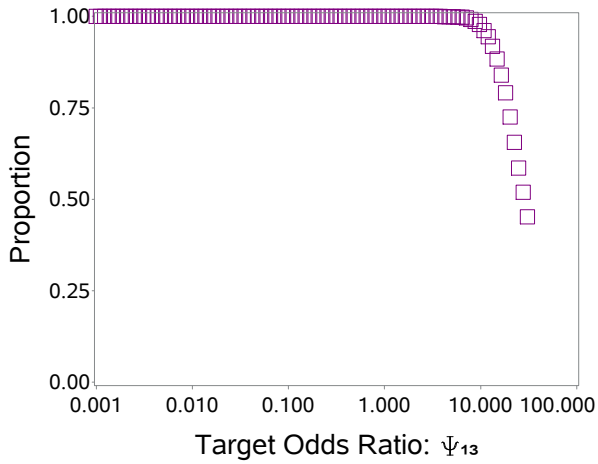


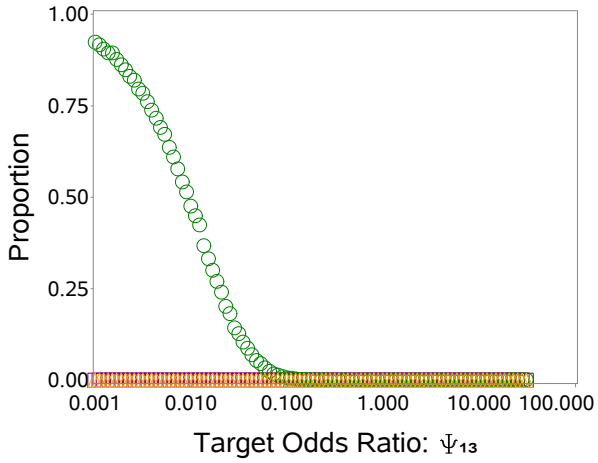
Fig. 3.10. Case: $p_1 = 0.6$, $p_2 = 0.5$, $p_3 = 0.6$; $\psi_{12} = 0.50$, $\psi_{23} = 1.75$ Proportion of Simulations with Odds Ratio Within the Fréchet Bounds

The proportion of simulations with the estimated odds ratio falling within the bounds stayed at 1 from the lower Fréchet bound of 0 until reaching target odds ratio 4.25, then it decreased to 0.452 at the upper bound of target odds ratio 30.86. See Figure 3.8.

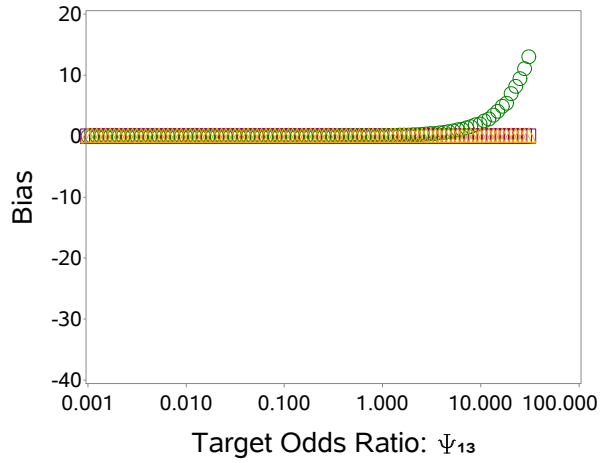
The proportion of simulations needing an adjustment to the odds ratio due to a zero cell was 0 for ψ_{12} and ψ_{23}. Starting at target odds ratio 0, this proportion for ψ_{13} was 0.924, and it dropped to 0 by target odds ratio 0.312, staying at 0 until reaching target 14.856, at which it began to increase slowly, reaching 0.001 at the upper Fréchet bound.

The bias was positive but fairly low for ψ_{12} (≈ 0.04) and ψ_{23} (≈ 0.19). For ψ_{13}, the bias increased fairly steadily from near 0 to 13.07 at the upper bound.

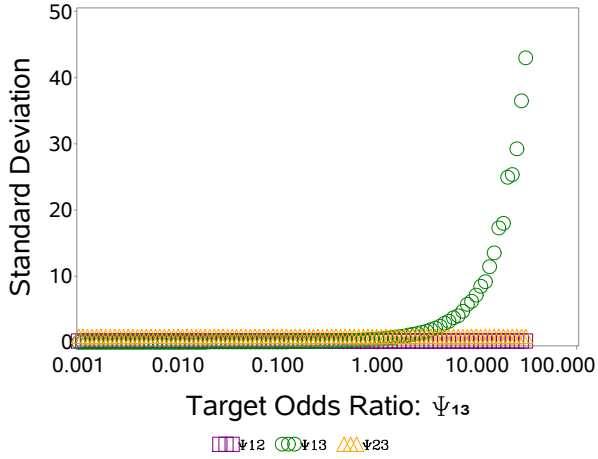
For ψ_{12}, the standard deviation stayed near 0.23 and for ψ_{23} it was near 0.89. For ψ_{13}, the



(a) Proportion of Simulations with Adjusted Odds Ratio



(b) Bias of Mean Odds Ratio Estimate



(c) Standard Deviation of Mean Odds Ratio Estimate

Fig. 3.11. Case: $p_1 = 0.6$, $p_2 = 0.5$, $p_3 = 0.6$; $\psi_{12} = 0.50$, $\psi_{23} = 1.75$ Plots for Measures of Interest

deviation increased from 0.003 starting at target odds ratio 0, up until 42.99 at the upper Fréchet bound.

See Figure 3.11 for plots of the various measures.

Remarks Regarding Unstructured Odds Ratio Case

While it would be unwise to make many generalizations based on only three cases, it is interesting to see the changes in the measures of interest of the estimates. In all cases, the proportion of simulations with adjusted odds ratio, the bias of the mean odds ratio, and the standard deviation of the mean odds ratio stayed mostly the same for ψ_{12} and ψ_{23} throughout the changes to ψ_{13}. The odds ratio is less restrictive to the simulation of variables than the correlation, therefore more variation in the simulated datasets is expected. Where the Fréchet bounds limited the odds ratios, the estimated odds ratios were likely to fall outside of the bounds when the target odds ratios were near the limits.

CHAPTER 4

FRÉCHET BOUNDS ON BINOMIAL AND NEGATIVE BINOMIAL DISTRIBUTIONS

4.1 Introduction

The Fréchet bounds have been studied for both binary and Poisson data, the descriptions of which can be found in a chapter written by Chaganty and Mav for a collection of works entitled *Computational Methods in Biomedical Research* [2]. Other distributions have not been thoroughly explored in terms of the Fréchet bounds. The binomial and negative binomial are common distributions encountered in discrete data and so will be investigated. Due to the combination functions found in both the binomial and negative binomial cumulative density functions (cdfs), there are no convenient closed form solutions for $E_L(y_i, y_j)$ and $E_U(y_i, y_j)$ as described in Equations 1.2 and 1.3. The bounds must therefore be calculated numerically.

4.2 Binomial Fréchet Bounds

The form of the right-tail cdf used to calculate the Fréchet bounds for the binomial distribution is

$$P(y_i \geq k) = \sum_{s=k}^{n_i} \binom{n_i}{s} p_i^s q_i^{n_i - s}$$

where n_i is the total number of trials for the random variable $\mathrm{Y_i}$ and q_i is $1 - p_i$. The Fréchet bounds then have the form

$$\rho_{ijL} = \frac{\sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \max \left[\sum_{s=k}^{n_i} \binom{n_i}{s} p_i^s q_i^{n_i - s} + \sum_{t=l}^{n_j} \binom{n_j}{t} p_j^t q_j^{n_j - t} - 1, 0 \right] - n_i p_i n_j p_j}{(n_i p_i q_i n_j p_j q_j)^{1/2}}$$

$$\rho_{ijU} = \frac{\sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \min \left[\sum_{s=k}^{n_i} \binom{n_i}{s} p_i^s q_i^{n_i - s}, \sum_{t=l}^{n_j} \binom{n_j}{t} p_j^t q_j^{n_j - t} \right] - n_i p_i n_j p_j}{(n_i p_i q_i n_j p_j q_j)^{1/2}}$$

as described in Equations 1.2 and 1.3.

4.2.1 Methods

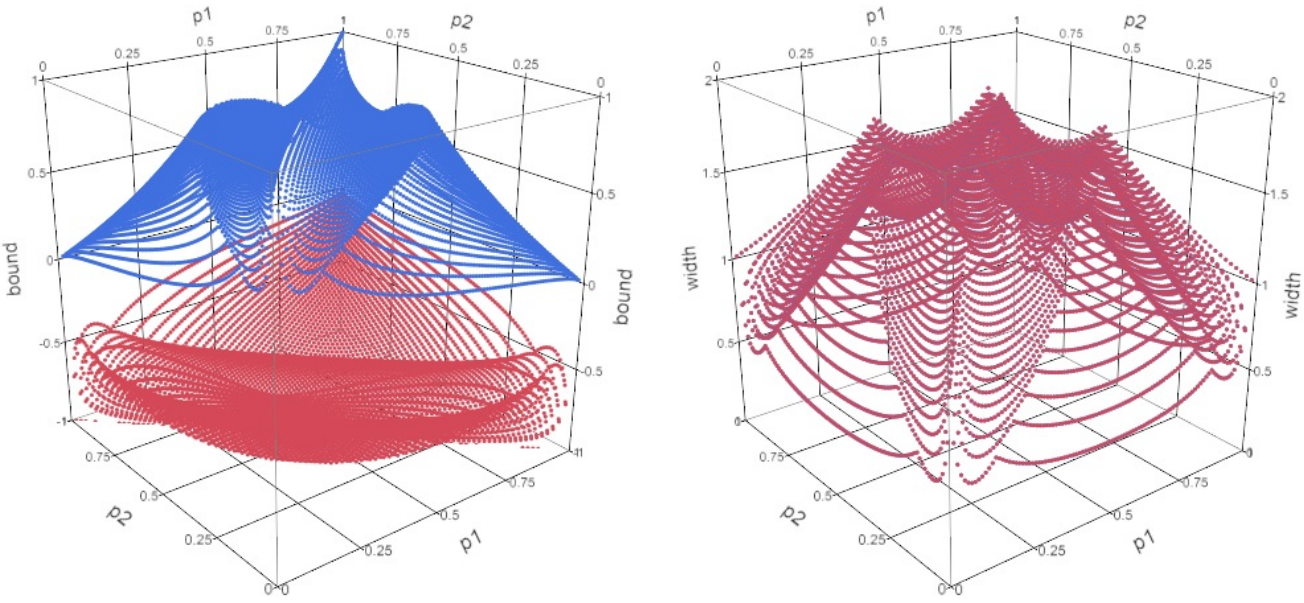
In order to obtain these bounds for various combinations of n_i, n_j, p_i, and p_j, a function for calculating the right-tail cdf of a binomial distribution was created. For each combination of the parameters chosen, the maxima and minima shown above were summed appropriately. After this, the bounds were calculated. Calculations were performed in SAS 9.4 (The SAS Institute, Cary, NC) using PROC IML. Figures were produced using JMP Pro 10.0.0 (The SAS Institute, Cary, NC).

4.2.2 Two-Variable Cases

As examples of the Fréchet bounds in two-variables cases, n_1 and n_2 were chosen from unique combinations of the integers 2, 10, 30, and 50. The bounds are shown for all sets of p_1 and p_2, starting at 0.01 and ending at 0.99 in increments of 0.01 for both probabilities. The widths of the intervals are presented, including the maximum, minimum, and median width with interquartile range (IQR). Recall that if width of the interval is narrow does not necessarily mean that there is little or no correlation between the variables, rather that there is a narrow range of what the correlation could possibly be.

Case: $n_1 = 2, n_2 = 2$

The Fréchet bounds in this case take on an interesting pattern when looked at as a whole in Figure 4.1a. The upper bounds take on a shape similar to the bottom of a boat, with the "keel" along the line $p_1 = p_2$. That is, the upper bound is at or near 1 along this line. Only along this line can the correlation between these two variables be perfect (i.e. $\rho_{12} = 1$). The lower bounds have the same overall shape, but "flipped over" and rotated 90 degrees, so that the keel is along the line $p_1 = 1 - p_2$. Only along this line can the two random variables be perfectly negatively correlated (i.e. $\rho_{12} = -1$).



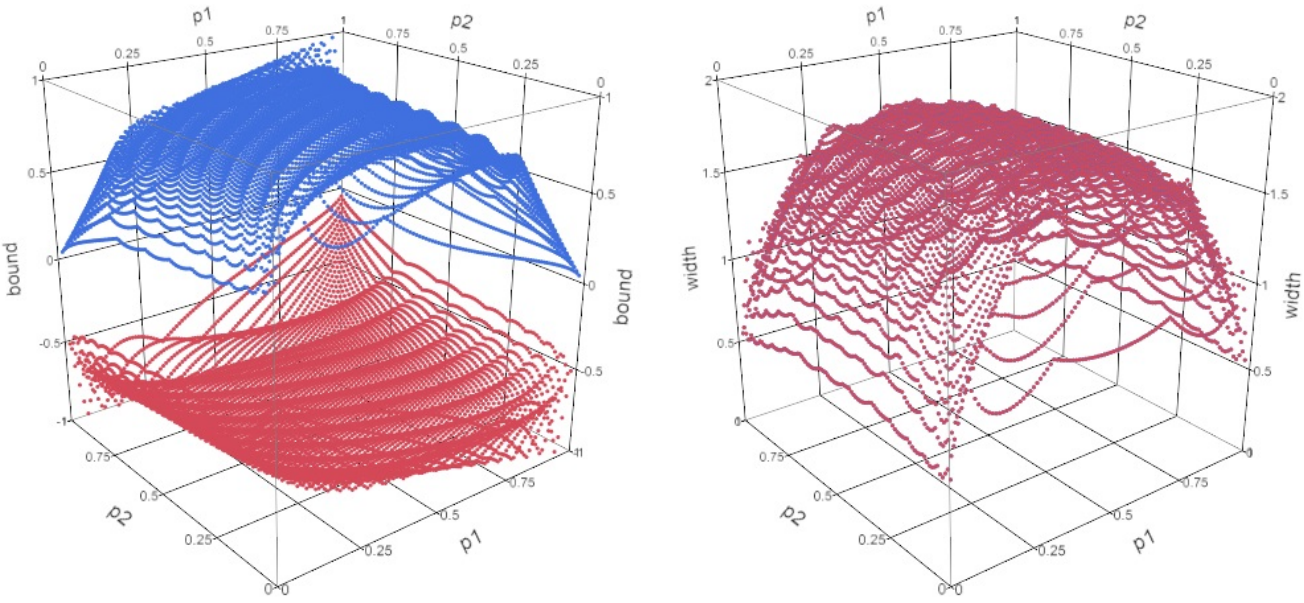
(a) Fréchet Bounds (b) Interval Widths

Fig. 4.1. Case: $n_1 = 2, n_2 = 2$

The widths of the intervals between the bounds are fairly narrow when either p_1 or p_2 are near 0 or 1, with a minimum width of 0.40. The median width is 1.35 with an IQR of [1.016, 1.529] There are nine distinct peaks in a symmetric pattern, with the maximum width at [0.5, 0.5, 2] ([p_1, p_2, width]). Other peaks are located at [0.14, 0.49 (0.51), 1.55], [0.29, 0.29, 1.82], [0.29, 0.71, 1.82], [0.49 (0.51), 0.14, 1.55], [0.49 (0.51), 0.86, 1.55], [0.71, 0.29, 1.82], [0.71, 0.71, 1.82], and [0.86, 0.49 (0.51), 1.55]. The widths have a tendency to get wider as p_1 and p_2 approach 0.5, giving the plot a bell-like shape. See Figure 4.1b. The peaks and troughs appear where they do according to the symmetry of the bounds, which correspond to the discrete nature of the distribution.

Case: $n_1 = 2, n_2 = 10$

The Fréchet bounds take on a ridged appearance in this case, with the ridges following an increasing slope as p_1 and p_2 increase for the upper bound, see Figure 4.2a. There appear to be ten ridges, corresponding to the 10 barriers "between" the 11 values in the support for Y_2 (i.e. l=0, 1,



(a) Fréchet Bounds (b) Interval Widths

Fig. 4.2. Case: $n_1 = 2, n_2 = 10$

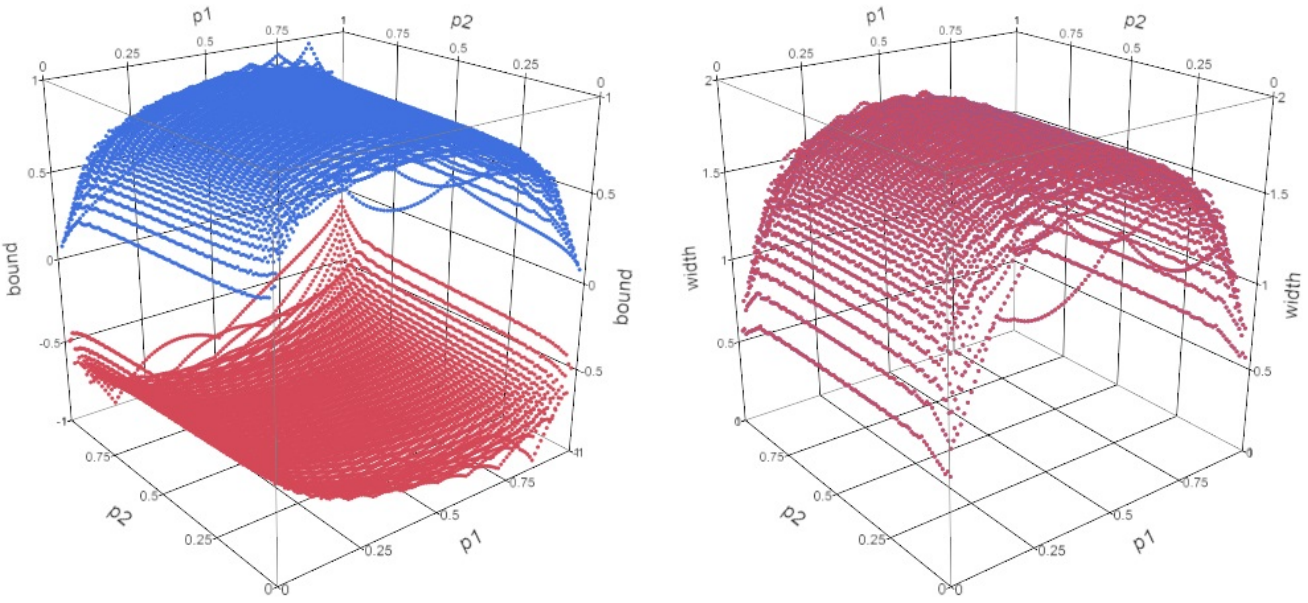
..., $n_2 = 10$). The upper bound approaches zero as p_1 approaches 1 and p_2 is near zero, as well as when p_2 approaches 1 where p_1 is near zero. The lower bound again takes on the same shape as the upper bound, but inverted and rotated as in the previous case.

The widths of the bounds (Figure 4.2b) are narrowest where p_1 is 0.01 or 0.99. The shape that the plot takes on has many small peaks and ridges, and it is somewhat of a twisted, parabolic, half-barrel shape. The minimum Fréchet interval width is 0.54, and the maximum is 1.83, which occurs where p_1 is 0.50 and p_2 is either 0.35 or 0.65. There are two maxima due to the 2 barriers between the 3 values in the support for Y_1. The median is 1.59 with IQR [1.356, 1.720].

Case: $n_1 = 2, n_2 = 30$

The Fréchet bounds in this case are nearly identical to the previous case, but with more and shallower ridges.

The widths of the bounds are also similar to the previous case, as would be expected. The minimum width is 0.57 and occurs at the "corners" of the plot, where p_1 and p_2 are both near 0.01



(a) Fréchet Bounds (b) Interval Widths

Fig. 4.3. Case: $n_1 = 2, n_2 = 30$

or 0.99. The maximum width is 1.80 and occurs where p_1 is 0.50 and p_2 is either 0.20 or 0.80. The median and IQR are 1.65 and [1.441, 1.748], respectively.

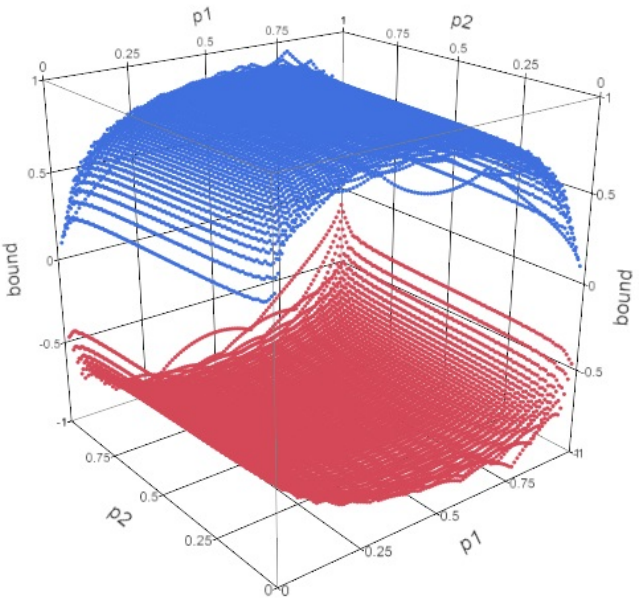
See Figure 4.3.

Case: $n_1 = 2, n_2 = 50$

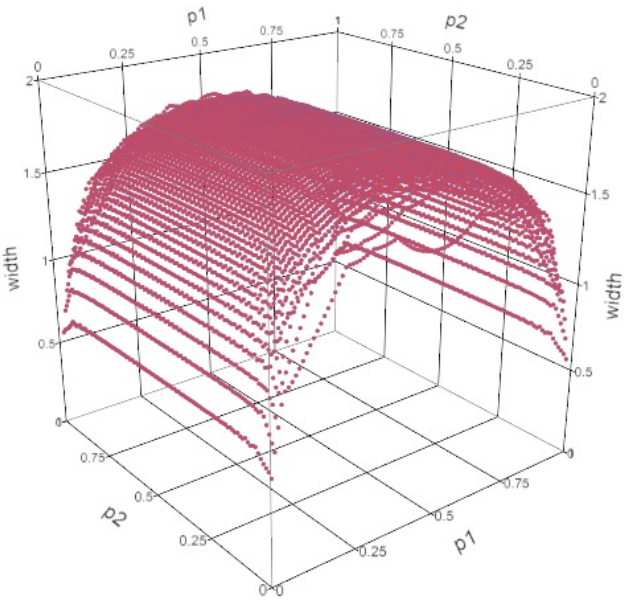
The Fréchet bounds in this case are nearly identical to the previous case, but with even more and shallower ridges.

The widths of the bounds are also similar to the previous case. The minimum width is 0.56 and occurs where p_1 and p_2 are both at 0.01 or 0.99. The maximum width is 1.80 and occurs where p_1 is 0.50 and p_2 is either 0.23 or 0.77. The median and IQR are 1.67 and [1.463, 1.758], respectively.

See Figure 4.4.

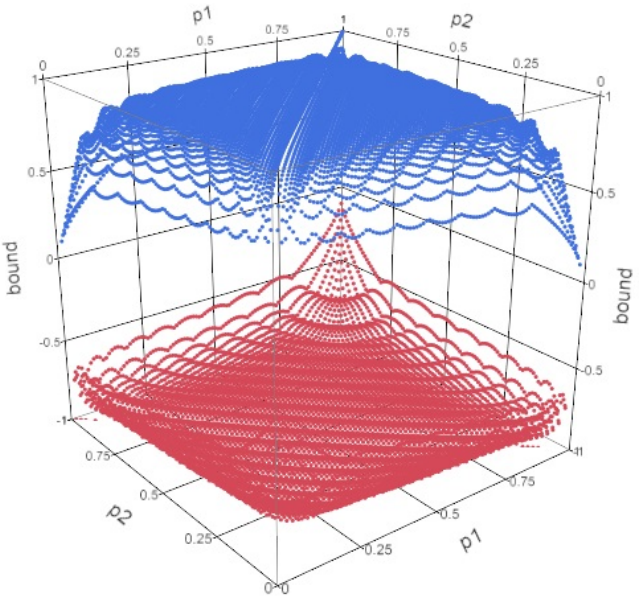


(a) Fréchet Bounds

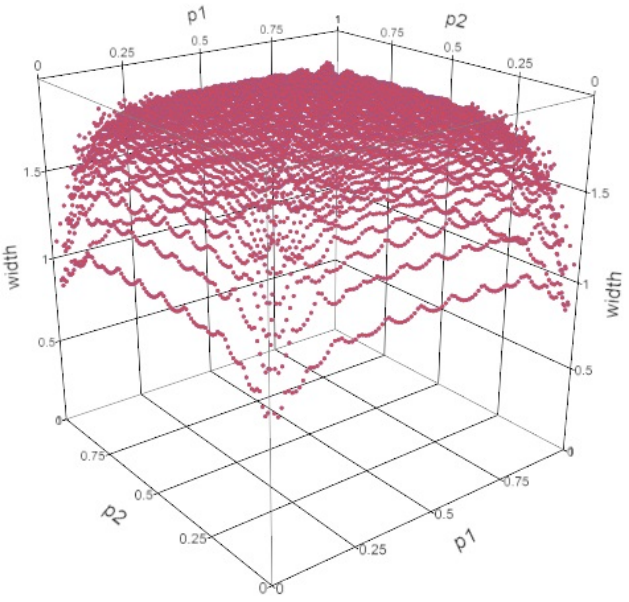


(b) Interval Widths

Fig. 4.4. Case: $n_1 = 2, n_2 = 50$



(a) Fréchet Bounds



(b) Interval Widths

Fig. 4.5. Case: $n_1 = 10, n_2 = 10$

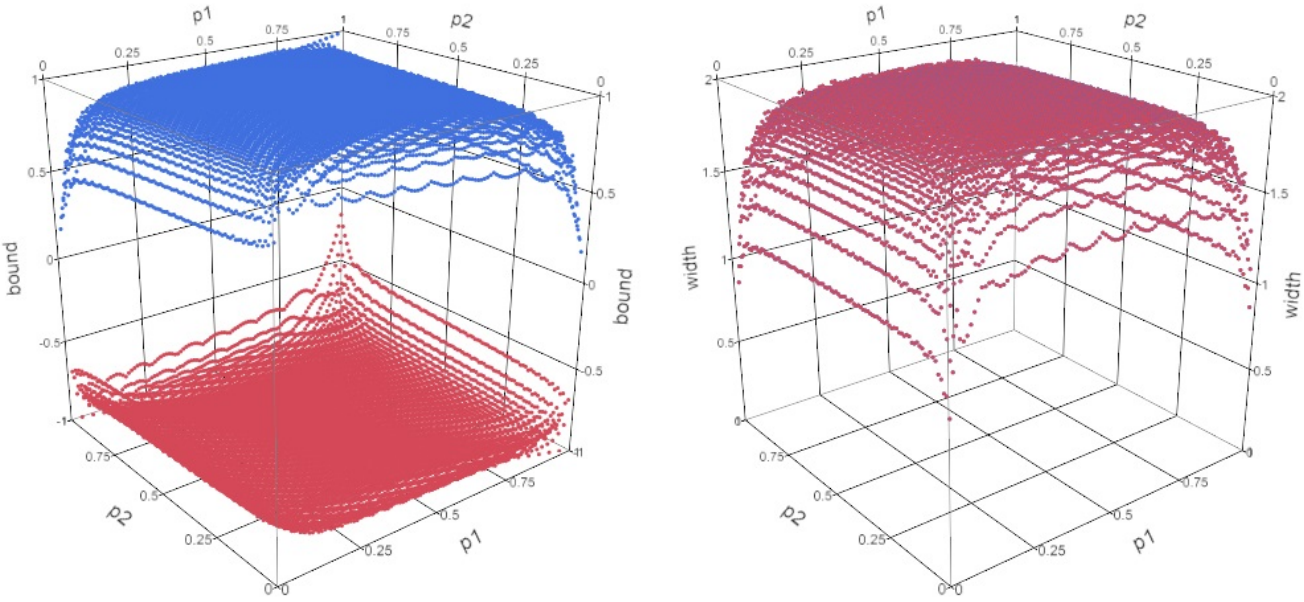
Case: $n_1 = 10, n_2 = 10$

The Fréchet bounds in this case again have symmetry as in the case where $n_1 = n_2 = 2$ along with a "keel" along the line $p_1 = p_2$ for the upper bound and $p_1 = 1 - p_2$ for the lower bound. Unlike the previously mentioned case, the bounds have something of an "egg crate" texture.

The minimum width is 0.85 and occurs where p_1 and p_2 are both near 0.01 or 0.99. The maximum width is 2.00 and occurs where both p_1 and p_2 are 0.50. The median and IQR are 1.85 and [1.714, 1.903], respectively.

See Figure 4.3.

Case: $n_1 = 10, n_2 = 30$



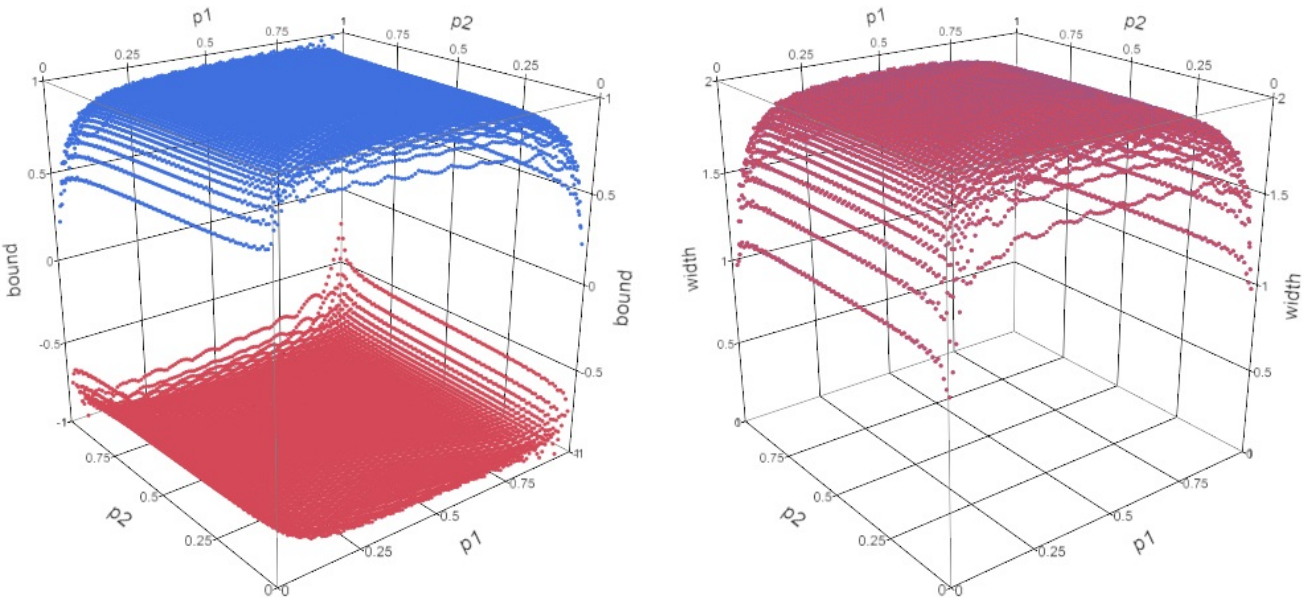
(a) Fréchet Bounds (b) Interval Widths

Fig. 4.6. Case: $n_1 = 10, n_2 = 30$

The "twisted barrel" shape reappears when $n_1 \neq n_2$. The shape of the bounds is flattening out toward 1 for the upper bounds and toward -1 for the lower bounds. The minimum width is 0.86 and occurs at the extreme corners, where p_1 and p_2 are both 0.01 or 0.99. The maximum width is

1.96 and occurs where p_1 = 0.5 and p_2 = 0.47 or 0.53. The median and IQR are 1.91 and [1.814, 1.941], respectively.

Case: $n_1 = 10, n_2 = 50$



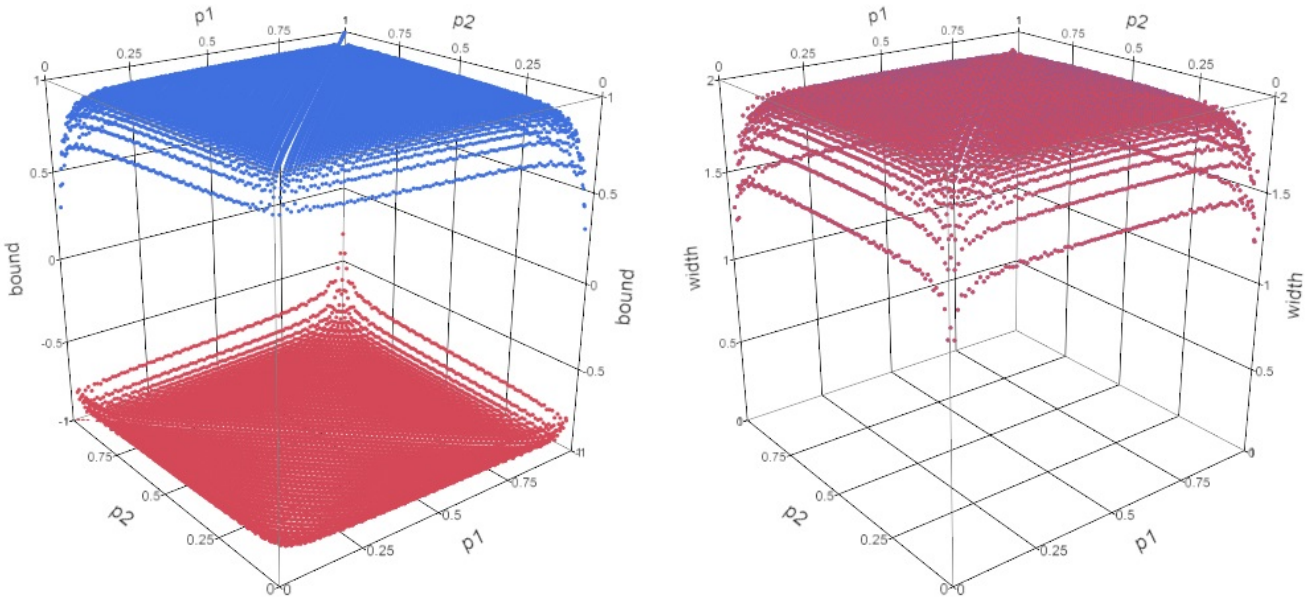
(a) Fréchet Bounds (b) Interval Widths

Fig. 4.7. Case: $n_1 = 10, n_2 = 50$

The shape of the bounds continues to flatten out toward 1 for the upper bounds and toward -1 for the lower bounds as n_2 increases. The minimum width is 0.97 and occurs at the extreme corners, where p_1 and p_2 are both 0.01 or 0.99. The maximum width is 1.97 and occurs where p_1 = 0.5 and p_2 = 0.29 or 0.71. The median and IQR are 1.92 and [1.836, 1.947], respectively.

Case: $n_1 = 30, n_2 = 30$

The shape of the bounds gains symmetry since $n_1 = n_2$, and the overall shape continues to flatten out toward 1 for the upper bounds and toward -1 for the lower bounds as both ns increase. The minimum width is 1.24 and occurs at the extreme corners, where p_1 and p_2 are both 0.01, 0.02, 0.98, or 0.99. The maximum width is 2.00 and occurs where $p_1 = p_2 = 0.5$. The median and



(a) Fréchet Bounds (b) Interval Widths

Fig. 4.8. Case: $n_1 = 30$, $n_2 = 30$

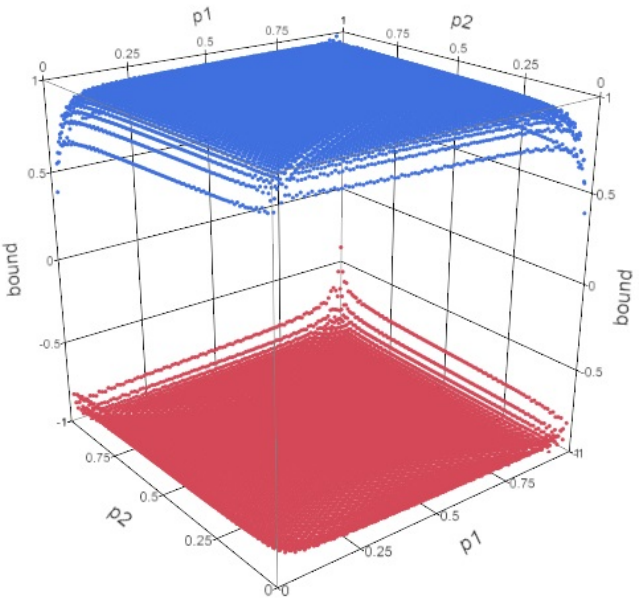
IQR are 1.95 and [1.908, 1.969], respectively.

Case: $n_1 = 30$, $n_2 = 50$

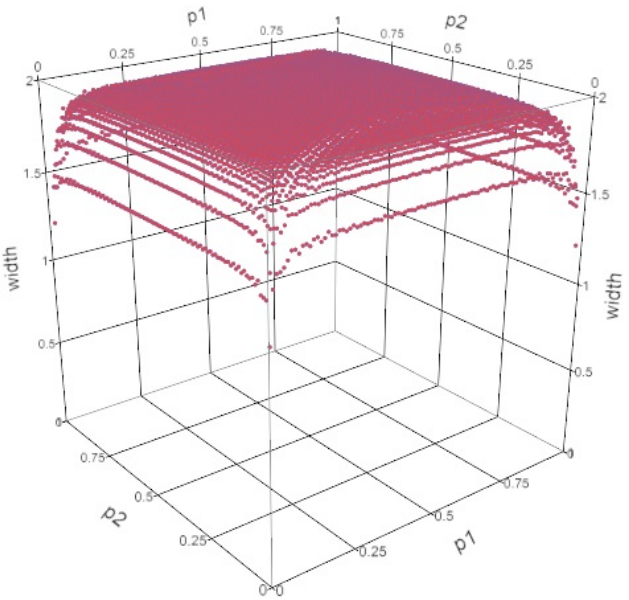
The shape of the bounds continues to flatten out toward 1 for the upper bounds and toward -1 for the lower bounds as the ns increase. The minimum width is 1.22 and occurs at the extreme corners, where p_1 and p_2 are both 0.01 or 0.99. The maximum width is 1.99 and occurs where p_1 = 0.45 or 0.55 and p_2 = 0.19 or 0.81. The median and IQR are 1.96 and [1.929, 1.977], respectively.

Case: $n_1 = 50$, $n_2 = 50$

The shape of the bounds gains symmetry since $n_1 = n_2$, and the overall shape continues to flatten out toward 1 for the upper bounds and toward -1 for the lower bounds as both ns increase. The minimum width is 1.51 and occurs at the extreme corners, where p_1 and p_2 are both 0.01 or 0.99. The maximum width is 2.00 and occurs where $p_1 = p_2 = 0.5$. The median and IQR are 1.97 and [1.947, 1.980], respectively.

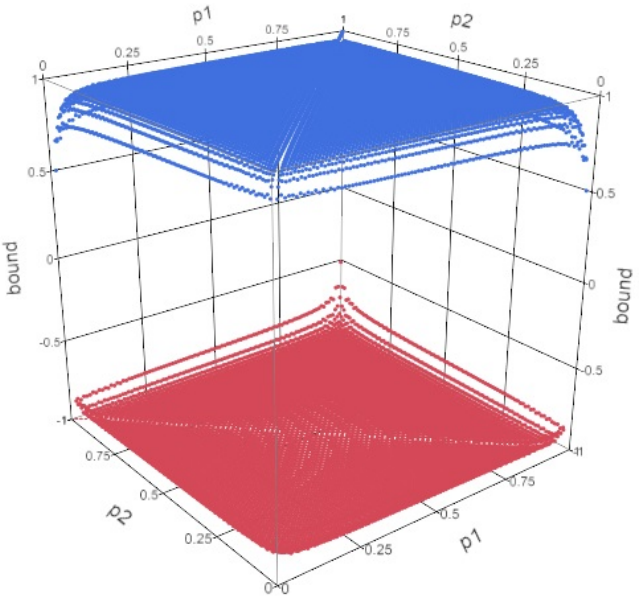


(a) Fréchet Bounds

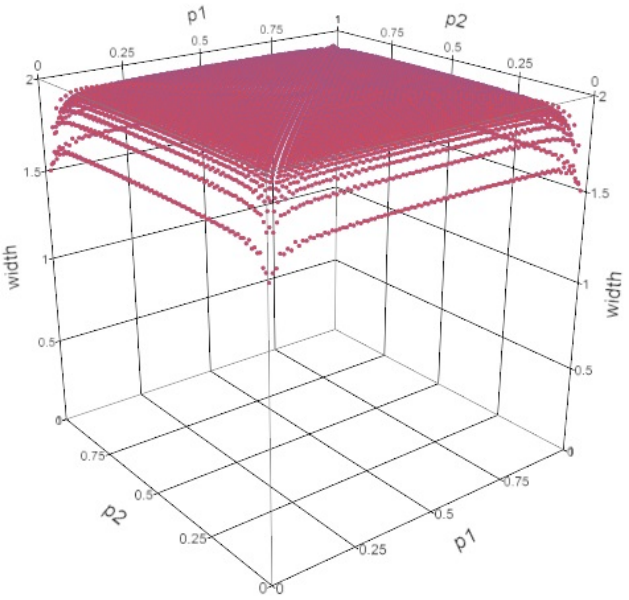


(b) Interval Widths

Fig. 4.9. Case: $n_1 = 30$, $n_2 = 50$



(a) Fréchet Bounds



(b) Interval Widths

Fig. 4.10. Case: $n_1 = 50$, $n_2 = 50$

4.2.2.1 Remarks on Two-Variable Cases

As n_1 and n_2 increase, so do the widths of the bounds. The upper bounds tend toward 1; the lower bounds tend toward -1, and the limitations on the correlation become less restrictive, though the "corners" remain more restrictive than the rest of the bounds in each case. This is to be expected. As n increases, most distributions tend to resemble the normal distribution. However, the only points at which the bounds are exactly [-1, 1] are when $n_1 = n_2$ and $p_1 = p_2 = 0.5$. The ridges seen in each figure appear to correspond to the n_i being used to calculate the bounds. For example, where The "twisted barrel" shape arises from discrepancy between n_1 and n_2.

4.2.3 Three-Variable Cases

In the three-variable cases, combinations of 2 and 10 were chosen for n_1, n_2, and n_3. The same grid as in the two-variable cases was used for p_1 and p_2 while p_3 varied between 0.25, 0.50, and 0.75.

The Fréchet bounds in the compound symmetric (CS) case are the most limiting of the bounds as calculated for each of the three correlations, since the CS structure assumes $\rho_{12} = \rho_{13} = \rho_{23} = \rho$. That is,

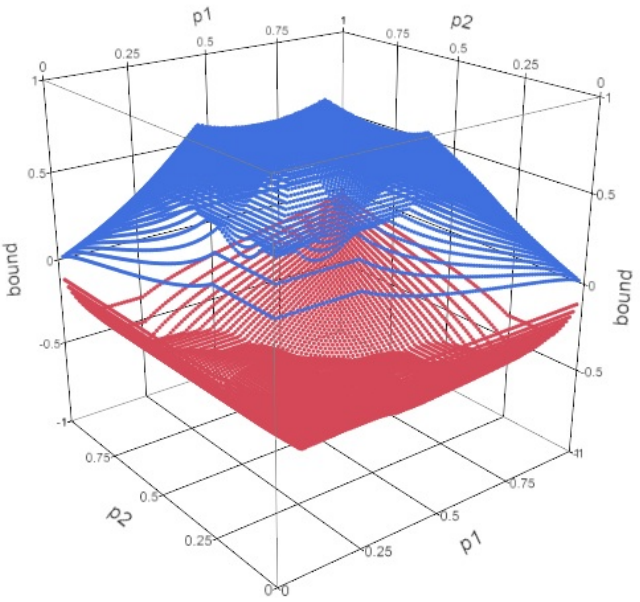
$$\max\left(\rho_{12L}, \rho^*_{13L}, \rho_{23L}\right) \leq \rho \leq \min\left(\rho_{12U}, \rho^*_{13U}, \rho_{23U}\right) \tag{4.1}$$

where $\rho^*_{13L} = \rho_{13L}$ and $\rho^*_{13U} = \rho_{13U}$.

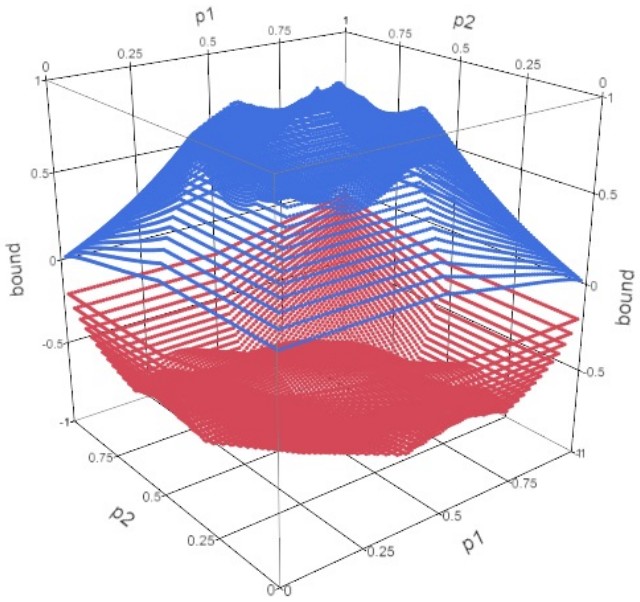
In the AR(1) case, the Fréchet bounds are slightly trickier since $\rho_{12} = \rho_{23} = \rho$ and $\rho_{13} = \rho^2$. Often in the AR(1) case, only positive correlations are considered due to the exponent on ρ_{13}, but the possibility of negative correlations will be allowed for the sake of completeness. In Equation (4.1), the quantities $\rho^*_{13L} = S(\rho_{13L})\sqrt{\rho_{13L}}$ and $\rho^*_{13U} = S(\rho_{13U})\sqrt{\rho_{13U}}$, where $S(\cdot)$ is the signum function.

Case (CS): $n_1 = 2, n_2 = 2, n_3 = 2$

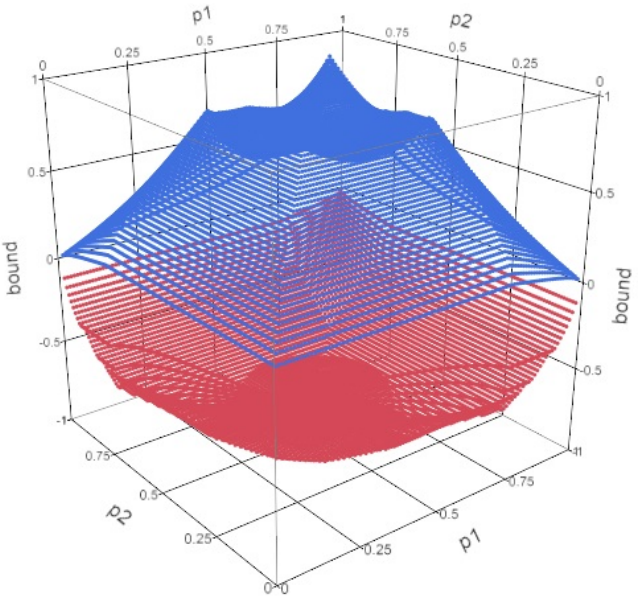
Though it may be difficult to see in Figure 4.11, the point of greatest magnitude for the upper bound shifts from (p_1, p_2)=(0.25, 0.25) to (0.50, 0.50) to (0.75, 0.75) as p_3 shifts from 0.25 to 0.50



(a) $p_3 = 0.25$



(b) $p_3 = 0.50$



(c) $p_3 = 0.75$

Fig. 4.11. Case (CS): $n_1 = 2, n_2 = 2, n_3 = 2$

to 0.75. The upper and lower bounds are symmetric about the plane $p_1 = p_2$ for each p_3.

There are four prominent peaks in the upper bounds of Figure 4.11a, at (0.25, 0.25), (0.25, 0.66), (0.66, 0.25), and (0.66, 0.66). The lower bounds take on a bowl-shaped appearance, with "feet" near the same points as the peaks of the upper bounds. The lower bounds do not always follow a relatively smooth curve as the upper bounds appear to.

There are seven prominent peaks in the upper bounds of Figure 4.11b, at (0.14, 0.14), (0.14, 0.50), (0.14, 0.86), (0.50, 0.50), (0.50, 0.86), and (0.86, 0.86). The lower bounds have sharp inverted peaks, with a minimum peak of -1 at (0.50, 0.50), and several shallower ones dispersed throughout, near the places where the peaks in the upper bounds are located. The lower bounds in this figure are not disjointed as the ones in Figure 4.11a.

There are four prominent peaks in the upper bounds of Figure 4.11c, at (0.34, 0.34), (0.34, 0.75), (0.75, 0.34), and (0.75, 0.75). The figure appears the same as Figure 4.11a, but it has been rotated.

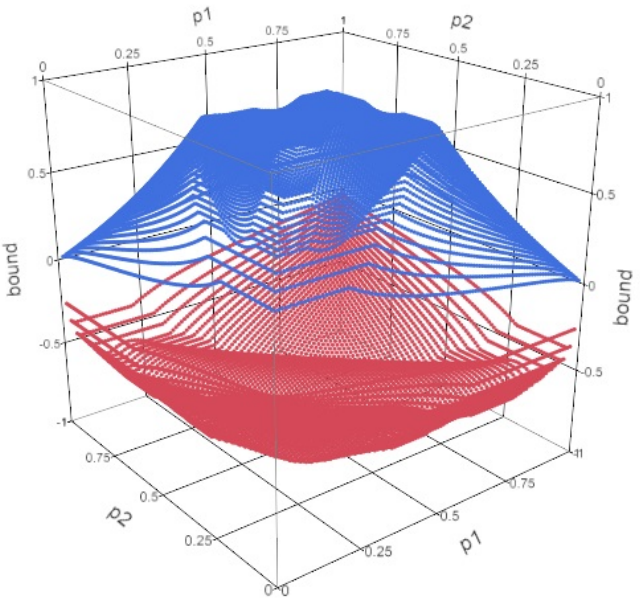
Case (CS): $n_1 = 2, n_2 = 2, n_3 = 10$

The upper bounds in this case appear more as three ridges rather than distinct peaks as in the previous section, as do the lower bounds.

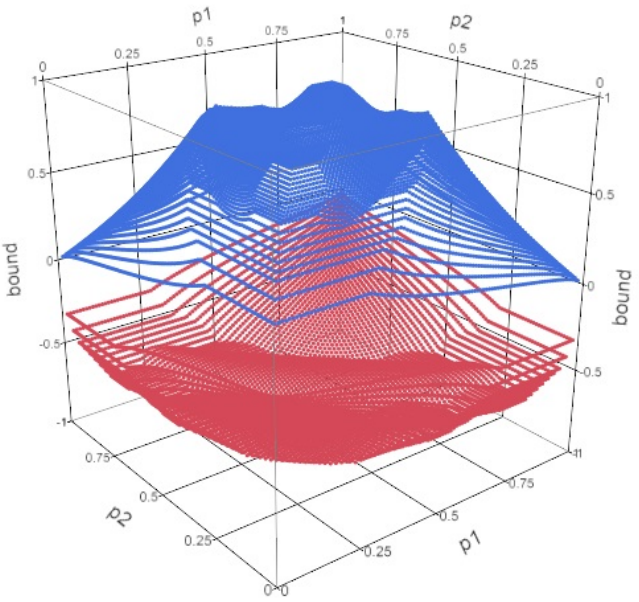
In Figure 4.12a, the middle ridge of the upper bounds has three major "bumps," which are at (0.28, 0.28), (0.48, 0.48), and (0.69, 0.69), with higher values near the first and second bumps. The other two ridges appear to curve smoothly around the middle ridge. The lower bounds are similar, but there is a seeming disjunction where one might expect the bounds to be smoothly connected at the corners.

In Figure 4.12b, the middle ridge of the upper bounds has two major bumps centering around (0.61, 0.61) and (0.39, 0.39).

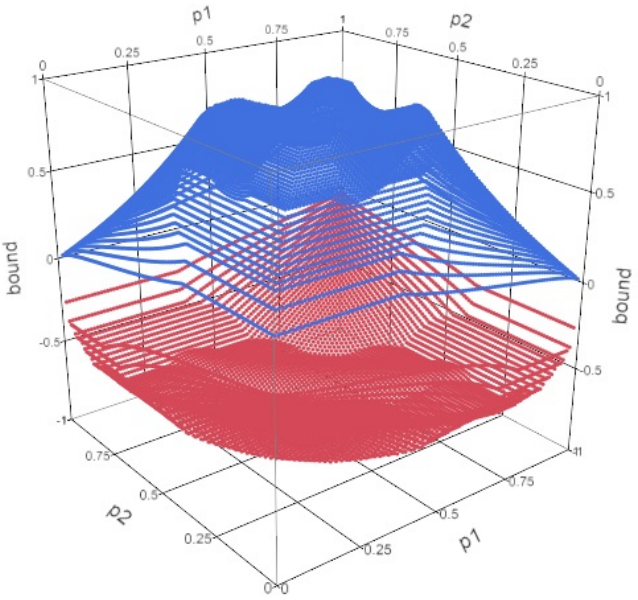
In Figure 4.12c, the middle ridge of the upper bounds has three major "bumps," which are at (0.31, 0.31), (0.52, 0.52), and (0.72, 0.72), with higher values near the second and third bumps. The other two ridges appear to curve smoothly around the middle ridge. This figure is symmetric



(a) $p_3 = 0.25$



(b) $p_3 = 0.50$



(c) $p_3 = 0.75$

Fig. 4.12. Case (CS): $n_1 = 2$, $n_2 = 2$, $n_3 = 10$

to Figure 4.12a.

Case (CS): $n_1 = 2, n_2 = 10, n_3 = 10$

In this case, the bounds take on a "twisted barrel" shape, as in the two-variable case where $n_1 \neq n_2$. The bounds also do not have smooth connections at the corners of the figures, as they do in the above three-variable cases.

As in the previous three-variable CS cases, 4.13c is a rotation of 4.13a, and 4.13b has some slight differences from the other two figures in the placement and depth of the ridges.

Case (CS): $n_1 = 10, n_2 = 10, n_3 = 10$

In these figures, the "egg crate" texture is seen as in 4.5a, but it appears shallower. As in previous three-variable cases, 4.14c is a rotation of 4.14a, and the maximum upper bound for each graph appears where $p_1 = p_2 = p_3$.

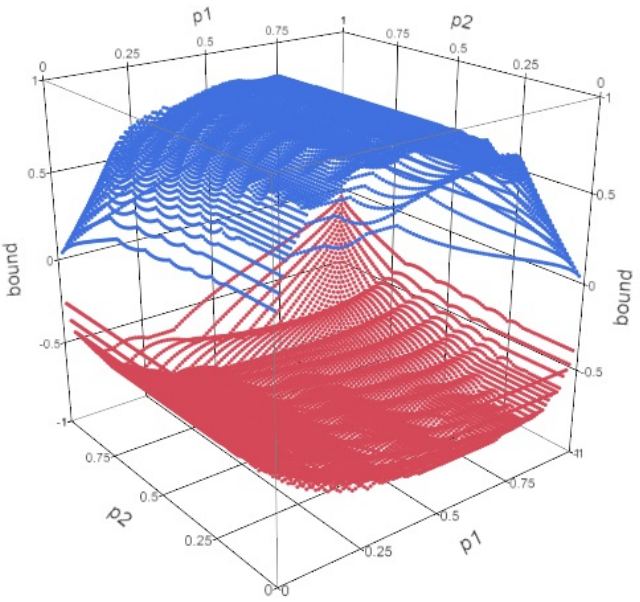
The minimum lower bound is at (0.65, 0.65) for Figure 4.14a, at (0.5, 0.5) for Figure 4.14b, and at (0.35, 0.35) for Figure 4.14c. Again, the lower bounds seem a little disrupted at the corners.

Case (AR(1)): $n_1 = 2, n_2 = 2, n_3 = 2$

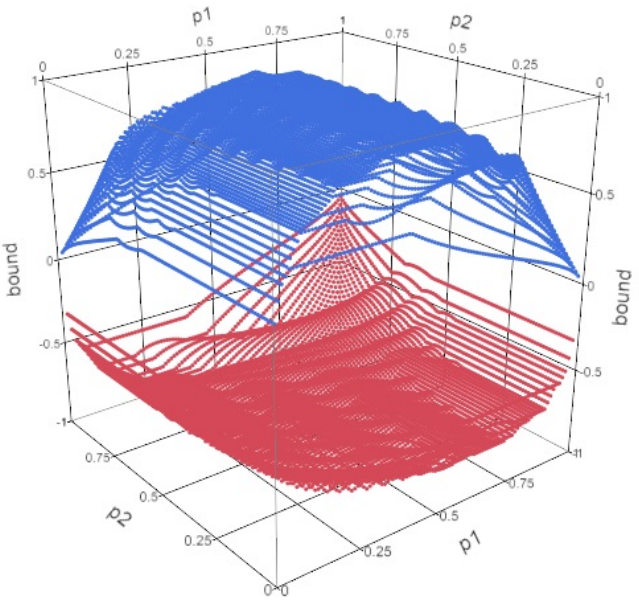
The AR(1) cases appear to be similar to the CS cases, though here they appear more twisted. The figures are all asymmetric, but there are similar ridges in the upper and lower bounds as in the CS cases. They are less smooth, appearing disjointed in the upper bounds where p_1 and p_2 are near the extreme lows and highs. Disjoints appear in the lower bounds where p_1 and p_2 are at opposite extremes (e.g. $p_1 = 0.99$ and $p_2 = 0.01$.

Case (AR(1)): $n_1 = 2, n_2 = 2, n_3 = 10$

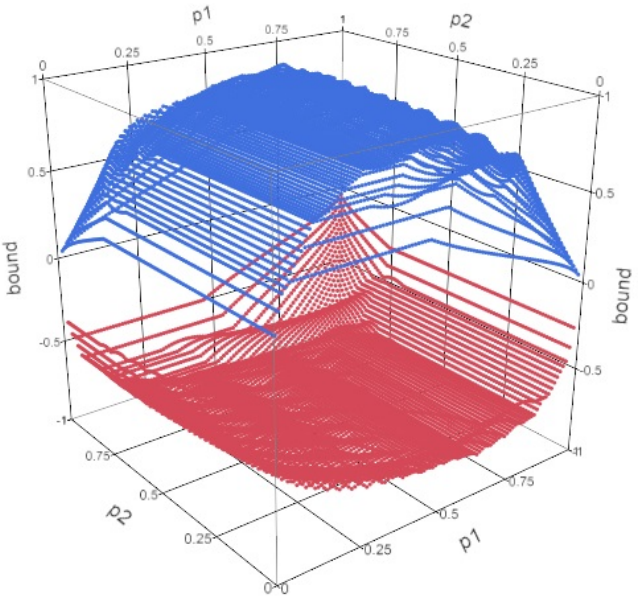
In this case, the "bumps" in the CS case of the same parameters have changed to slightly different shapes, but they are in similar location. There are also more disjoints than in the CS case. That is, the bounds do not match up in a smooth manner. Compare Figure 4.16 to Figure 4.12, paying attention to the corners.



(a) $p_3 = 0.25$

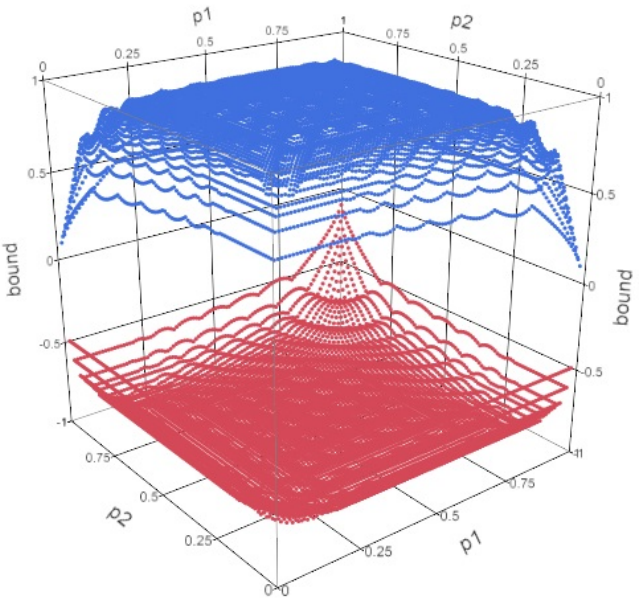


(b) $p_3 = 0.50$

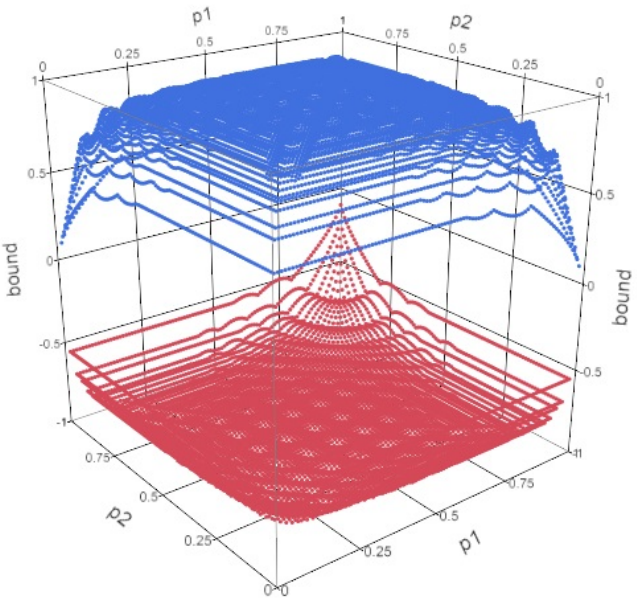


(c) $p_3 = 0.75$

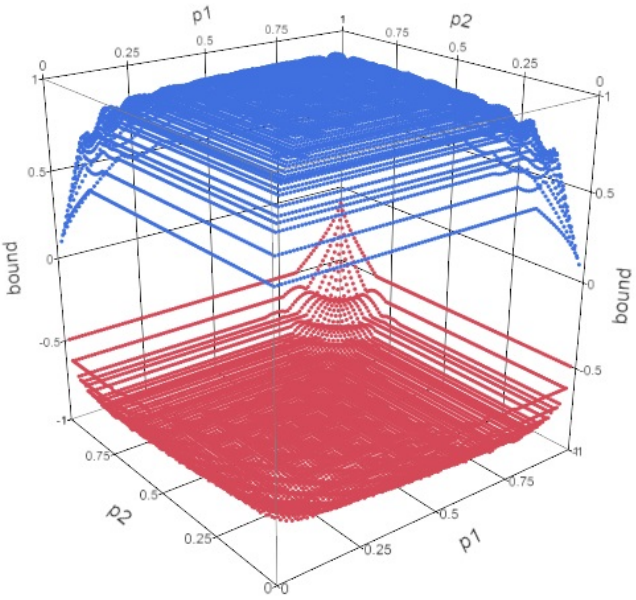
Fig. 4.13. Case (CS): $n_1 = 2$, $n_2 = 10$, $n_3 = 10$



(a) $p_3 = 0.25$

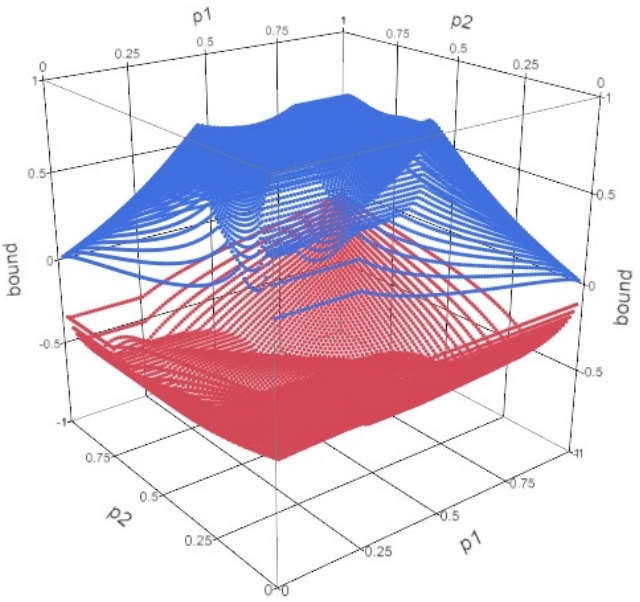


(b) $p_3 = 0.50$

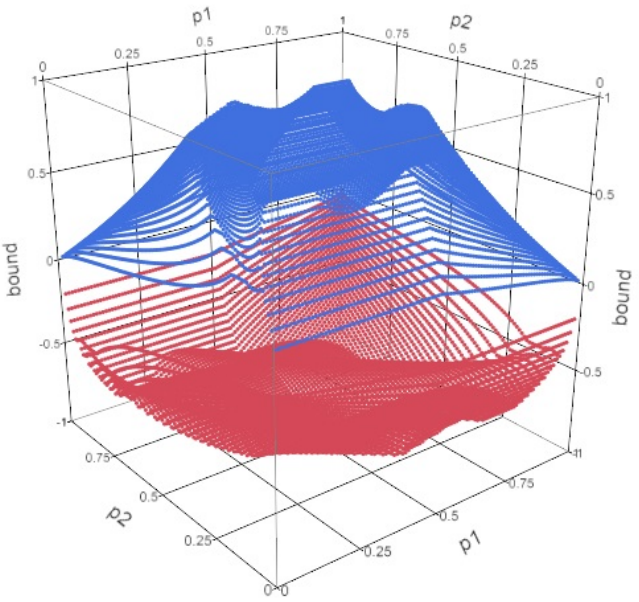


(c) $p_3 = 0.75$

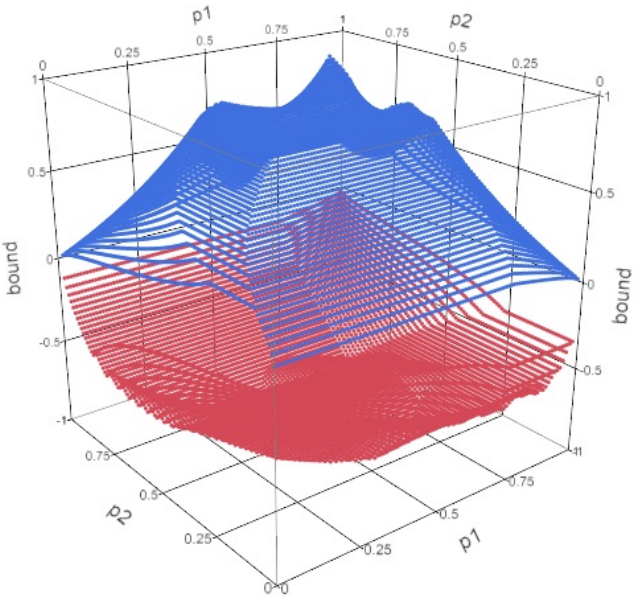
Fig. 4.14. Case (CS): $n_1 = 10$, $n_2 = 10$, $n_3 = 10$



(a) $p_3 = 0.25$

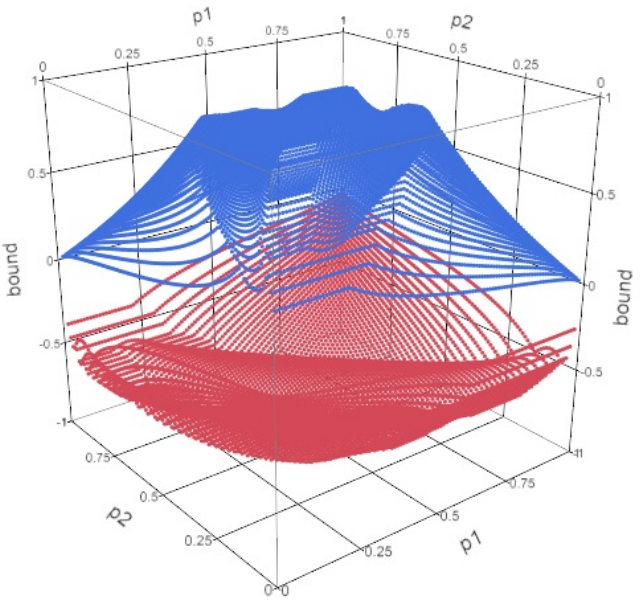


(b) $p_3 = 0.50$

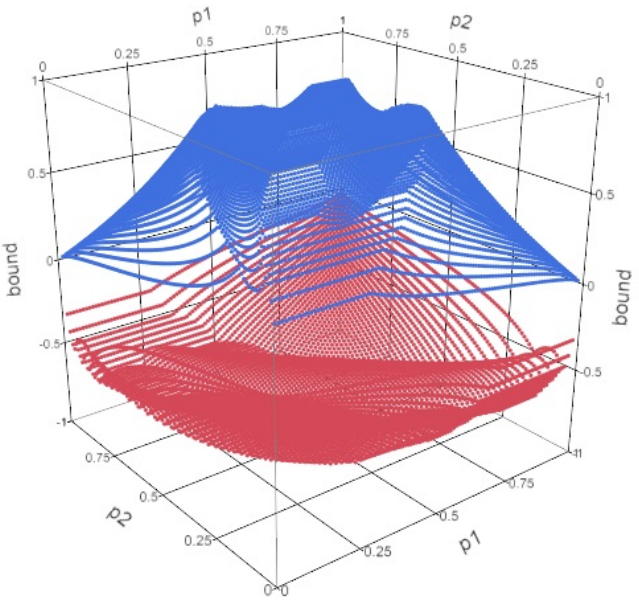


(c) $p_3 = 0.75$

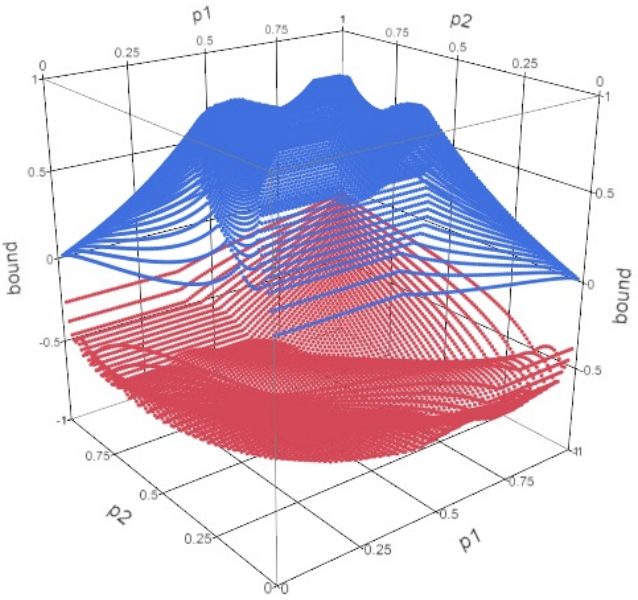
Fig. 4.15. Case (AR(1)): $n_1 = 2$, $n_2 = 2$, $n_3 = 2$



(a) $p_3 = 0.25$



(b) $p_3 = 0.50$



(c) $p_3 = 0.75$

Fig. 4.16. Case (AR(1)): $n_1 = 2, n_2 = 2, n_3 = 10$

Case (AR(1)): $n_1 = 2, n_2 = 10, n_3 = 10$

Unlike the previous cases, in this case the figures appear smoother and less disjointed than in the CS case of the same parameters. The same twisted, off-center shape is present as in Figure 4.13

Case (AR(1)): $n_1 = 10, n_2 = 10, n_3 = 10$

The bounds in this case are again more disjointed as compared to the CS case of the same parameters, but with the same egg crate appearance. The maxima are in the same locations, that is, where $p_1 = p_2 = p_3$.

Remarks Regarding Three-Variable Cases

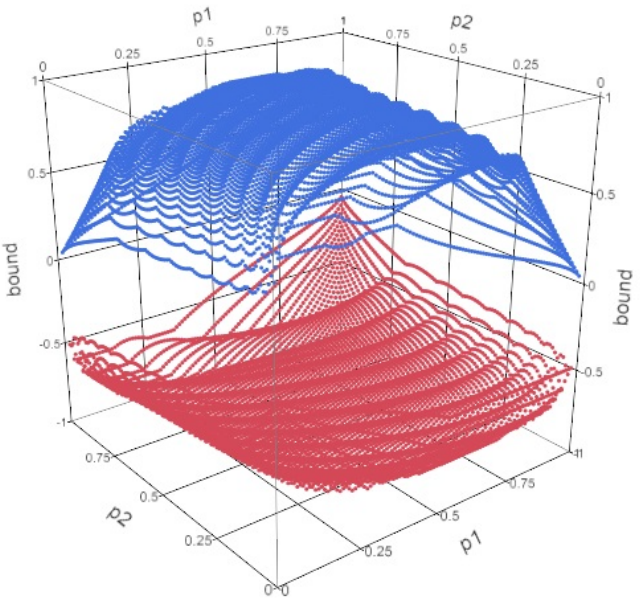
For both the CS and AR(1) cases, as the n_i increase, the Fréchet bounds become wider and have more ridges and peaks. Also, if $n_1 \neq n_2$, the figures take on a twisted appearance. The CS cases appear symmetric along the plane $p_1 = p_2$ if $n_1 = n_2$ regardless of n_3. The AR(1) cases follow the same general appearance as the CS cases, however, they do not appear smooth and are disjointed and asymmetric even in the cases where $n_1 = n_2 = n_3$. The changes in p_3 create obvious shifts in the upper and lower bounds.

4.3 Negative Binomial Fréchet Bounds

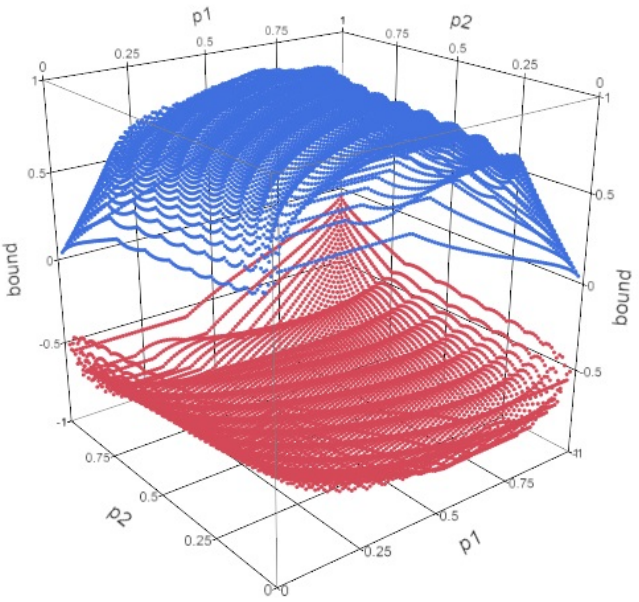
The form of the right-tail cdf used to calculate the Fréchet bounds for the negative binomial distribution is

$$P(y_i \geq k) = 1 - \sum_{s=0}^{k-1} \binom{s+r-1}{s} p_i^{r_i} (1-p_i)^s$$

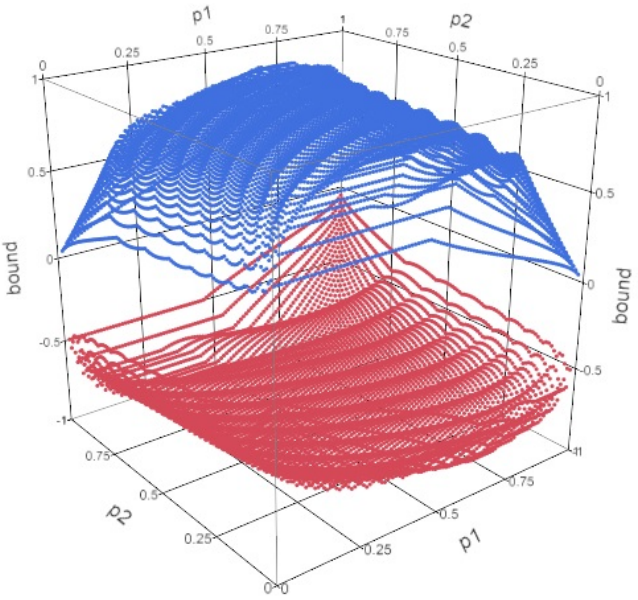
where s is the number of successes before the r_ith failure, p_i is the probability of failure, and q_i is $1 - p_i$.



(a) $p_3 = 0.25$

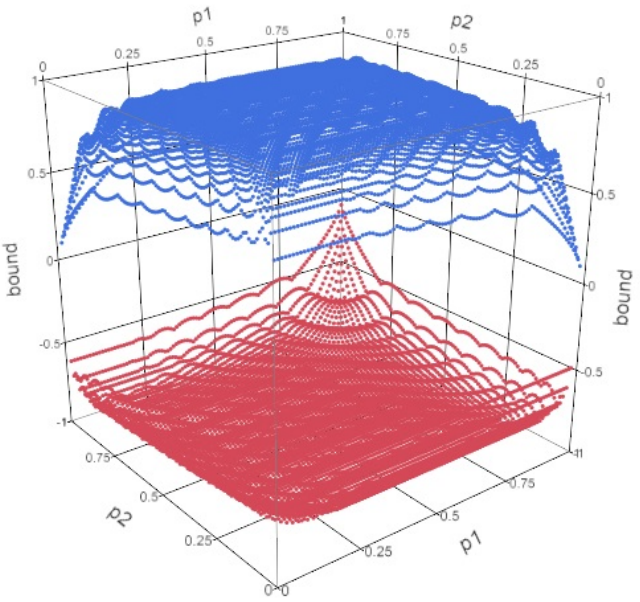


(b) $p_3 = 0.50$

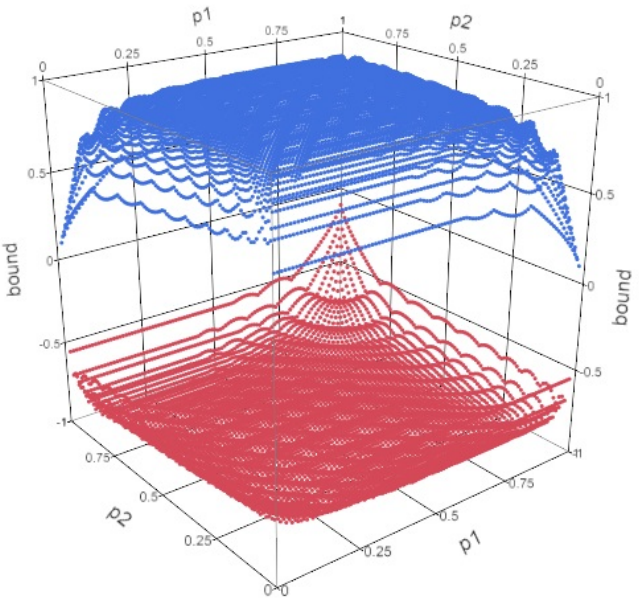


(c) $p_3 = 0.75$

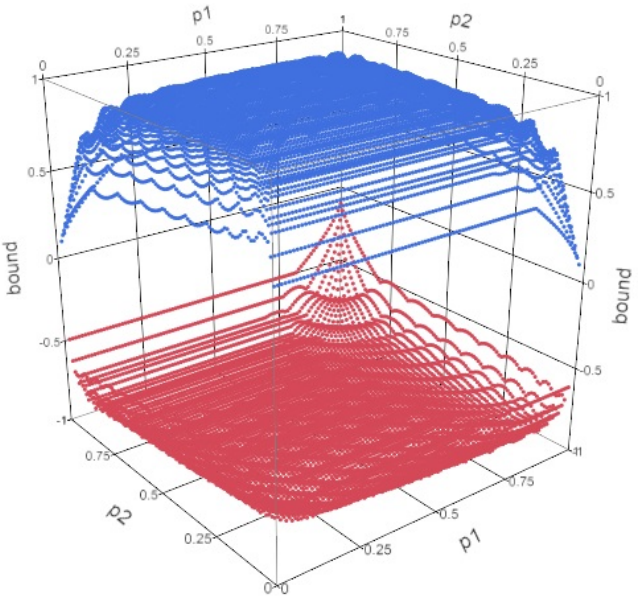
Fig. 4.17. Case (AR(1)): $n_1 = 2$, $n_2 = 10$, $n_3 = 10$



(a) $p_3 = 0.25$



(b) $p_3 = 0.50$



(c) $p_3 = 0.75$

Fig. 4.18. Case (AR(1)): $n_1 = 10$, $n_2 = 10$, $n_3 = 10$

The Fréchet bounds then have the form

$$\rho_{ijL} = \frac{E_{ijL} - \frac{r_i q_i r_j q_j}{p_i p_j}}{\left(\frac{r_i q_i r_j q_j}{p_i^2 p_j^2}\right)^{1/2}}$$

$$\rho_{ijU} = \frac{E_{ijU} - \frac{r_i q_i r_j q_j}{p_i p_j}}{\left(\frac{r_i q_i r_j q_j}{p_i^2 p_j^2}\right)^{1/2}}$$

where

$$E_{ijL} = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \max\left[1 - \sum_{s=0}^{k-1}\binom{s+r_i-1}{s} p_i^{r_i} q_i^s - \sum_{t=0}^{l-1}\binom{t+r_j-1}{t} p_j^{r_j} q_j^t, 0\right]$$

$$E_{ijU} = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \min\left[1 - \sum_{s=0}^{k-1}\binom{s+r_i-1}{s} p_i^{r_i} q_i^s, 1 - \sum_{t=0}^{l-1}\binom{t+r_j-1}{t} p_j^{r_j} q_j^t\right]$$

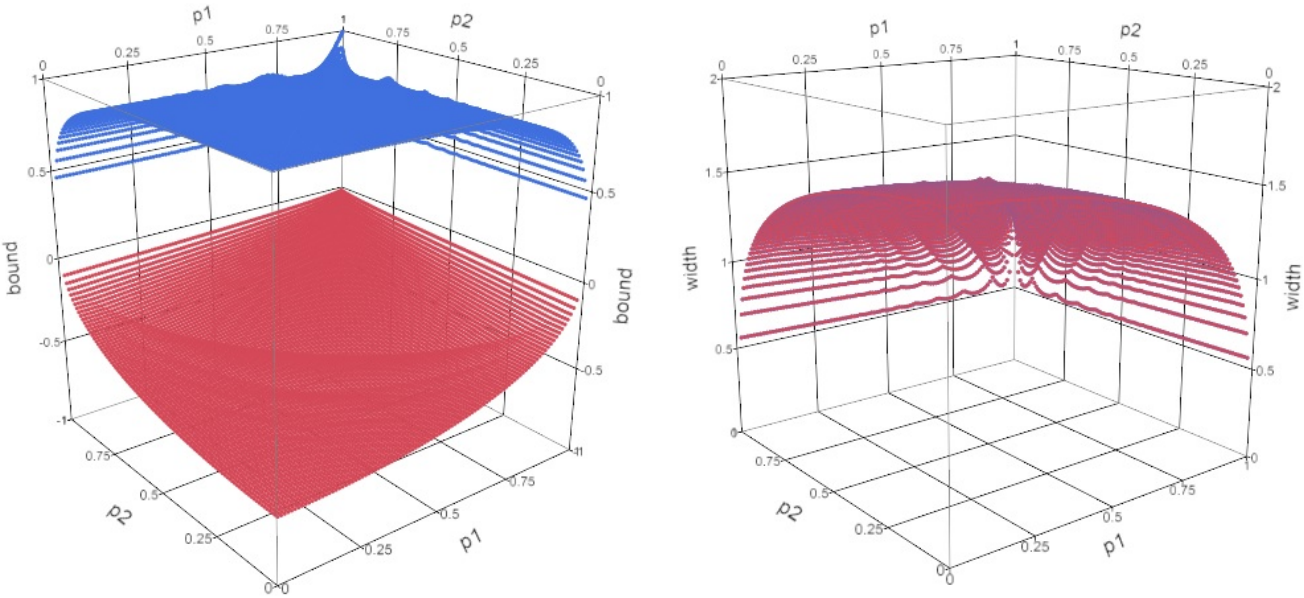
as described in Equations 1.2 and 1.3.

In order to obtain these bounds for various combinations of r_i, r_j, p_i, and p_j, a function for calculating the right-tail cdf of a negative binomial distribution was created. This was done by calculating the left-tail cdf and subtracting it from 1. The parameters k and l were chosen to be equal and increased iteratively by 1 according to the following criteria. For each combination of the parameters chosen, vectors allowing for the combinations of s and t corresponding to the maxima and minima above were created and then summed with the appropriate results of the previous iteration in order to find E_{ijL} and E_{ijU}. Using these quantities, the correlation bounds were calculated and compared to the previous iteration. If both the upper and lower bounds were within the general correlation bounds [-1, 1], the lower bound less than or equal to the upper bound, and within 0.000001 of the previous iteration, the algorithm was stopped and the Fréchet bounds were assumed to be found. Otherwise, the algorithm was repeated. Calculations were performed in SAS 9.4 (The SAS Institute, Cary, NC) using PROC IML. Figures were produced using JMP Pro 10.0.0 (The SAS Institute, Cary, NC).

4.3.1 Two-Variable Cases

As examples of the Fréchet bounds in two-variables cases, r_1 and r_2 were chosen from unique combinations of the integers 1, 2, 4, and 10. The bounds are shown for all sets of p_1 and p_2, starting at 0.01 and ending at 0.99 in increments of 0.01 for both probabilities. The widths of the intervals are presented, including the maximum, minimum, and median width with interquartile range (IQR).

Case: $r_1 = 1, r_2 = 1$



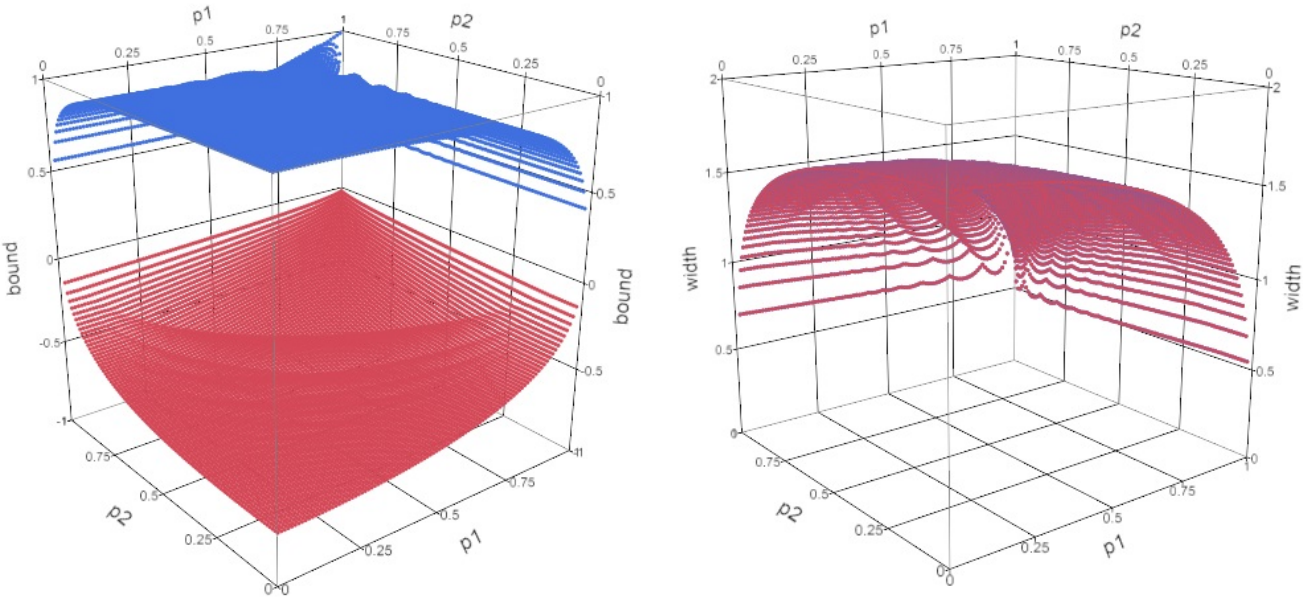
(a) Fréchet Bounds (b) Interval Widths

Fig. 4.19. Case: $r_1 = 1, r_2 = 1$

The Fréchet bounds were symmetric along the plane $p_1 = p_2$, with the upper bounds having ridges running parallel to the same plane and a "dorsal fin" similar to the keel of the binomial bounds. The lower bounds curve gently from the minimum to nearly zero along the same plane, with ridges running across the plane $p_1 = p_2$. The upper bound had a maximum of 1 at (0.99, 0.99) and the lower bound had a minimum of -0.64 at (0.01, 0.01). The width had a maximum of 1.64

and a minimum of 0.53, with median and IQR of 1.36 and [1.120, 1.522], respectively.

Case: $r_1 = 1, r_2 = 2$



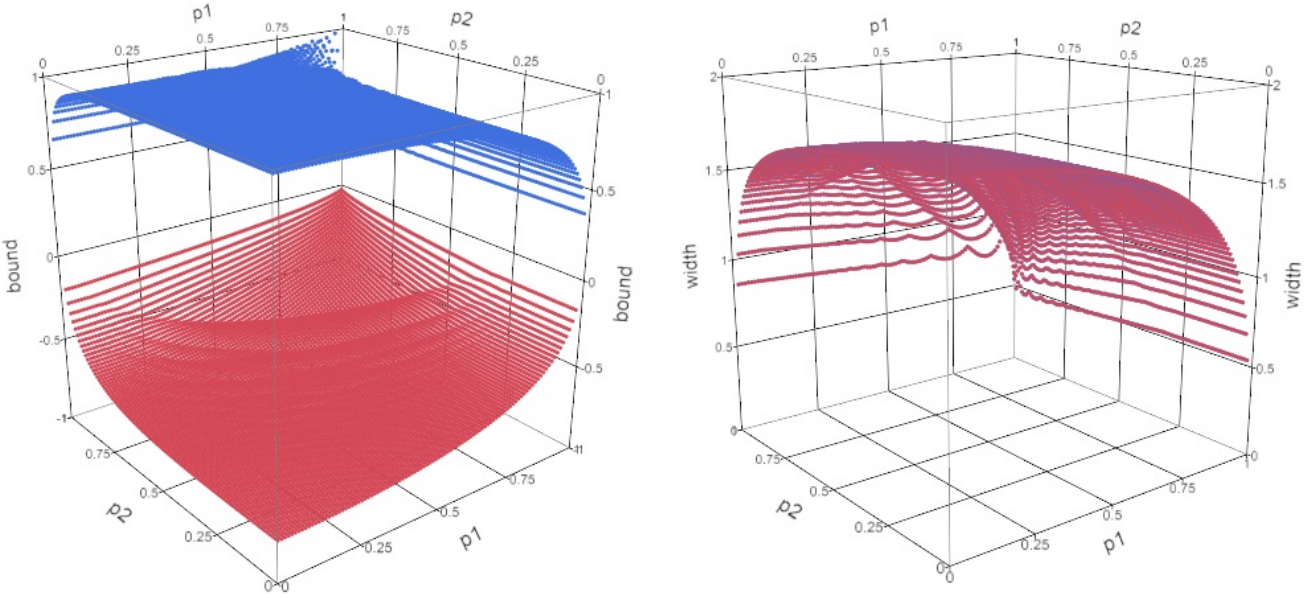
(a) Fréchet Bounds (b) Interval Widths

Fig. 4.20. Case: $r_1 = 1, r_2 = 2$

The Fréchet bounds are not symmetric in this case, but the upper bounds have similar ridges to the previous case, though shifted off center. The dorsal fin structure can be seen, though it is also off center. The lower bounds curve from the minimum to nearly zero, with ridges running across and also shifted. The upper bound has a maximum of 1.00 (0.995) at (0.98, 0.99) and the lower bound has a minimum of -0.73 at (0.01, 0.01). The width has a maximum of 1.72 and a minimum of 0.49, with median and IQR of 1.48 and [1.227, 1.625], respectively.

Case: $r_1 = 1, r_2 = 4$

The structure here is similar to that of the previous case, though the ridges are shallower and the shift is greater. The upper bound had a maximum of 0.99 at (0.96, 0.99) and the lower bound had a minimum of -0.78 at (0.01, 0.01). The width had a maximum of 1.76 and a minimum of



(a) Fréchet Bounds (b) Interval Widths

Fig. 4.21. Case: $r_1 = 1$, $r_2 = 4$

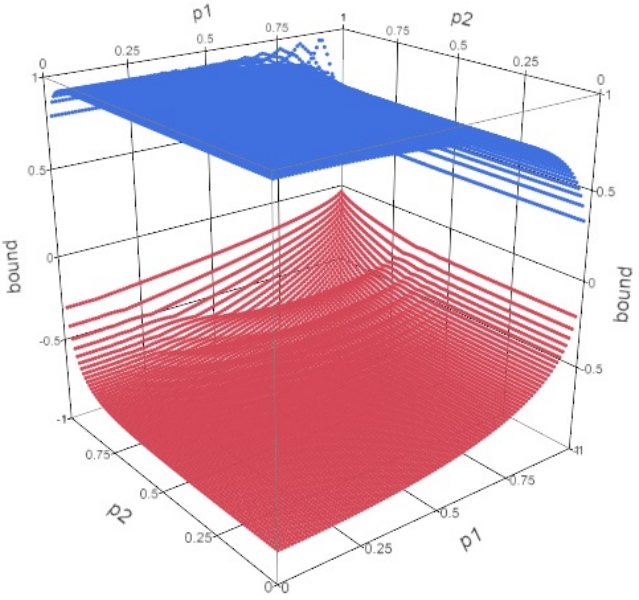
0.47, with median and IQR of 1.57 and [1.324, 1.691], respectively.

Case: $r_1 = 1$, $r_2 = 10$

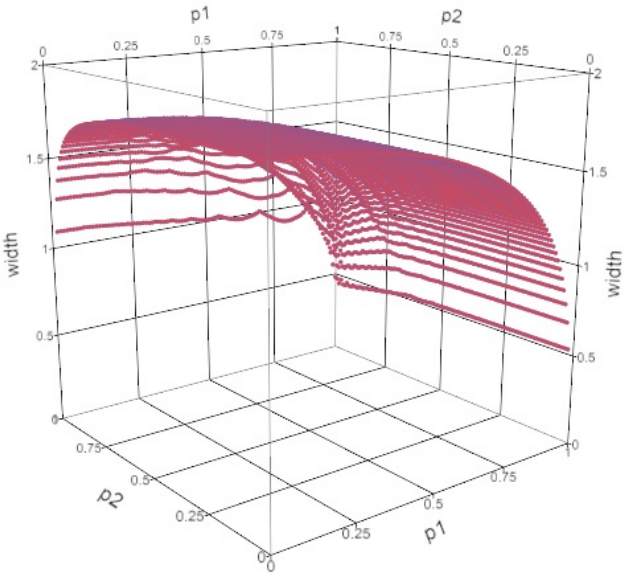
The structure here is similar to that of the previous case, though the ridges are even shallower and the shift is even greater. The upper bound had a maximum of 0.96 at (0.82, 0.99) and the lower bound had a minimum of -0.83 at (0.01, 0.01). The width had a maximum of 1.79 and a minimum of 0.45, with median and IQR of 1.63 and [1.411, 1.739], respectively.

Case: $r_1 = 2$, $r_2 = 2$

The Fréchet bounds are symmetric again, as might be expected from the binomial examples. The upper bounds and lower bounds are generally further apart than in the case where $r_1 = 1$ and $r_2 = 1$, making for wider intervals, however the shape remains about the same. The upper bound had a maximum of 1 at (0.99, 0.99) and the lower bound had a minimum of -0.80 at (0.01, 0.01). The width had a maximum of 1.80 and a minimum of 0.60, with median and IQR of 1.62 and

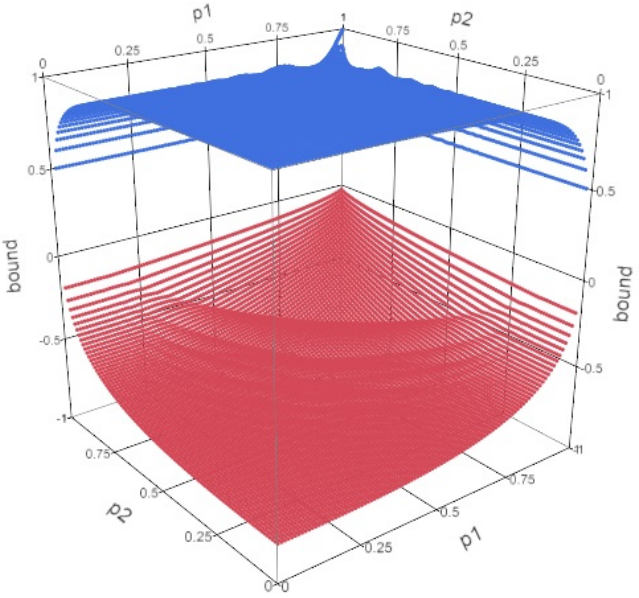


(a) Fréchet Bounds

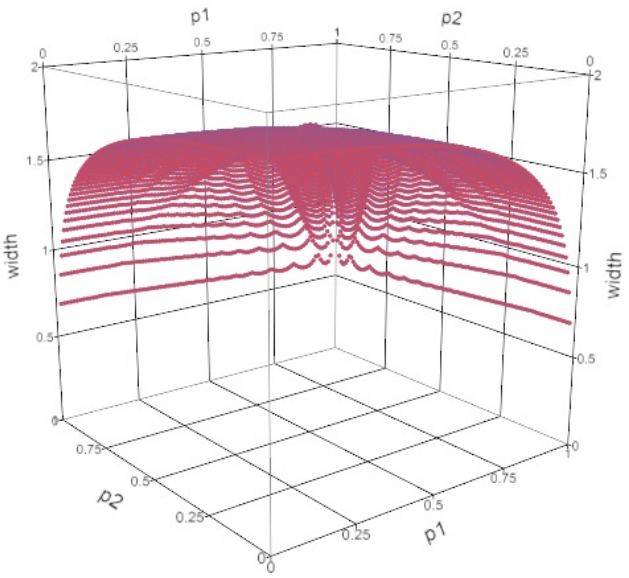


(b) Interval Widths

Fig. 4.22. Case: $r_1 = 1$, $r_2 = 10$



(a) Fréchet Bounds

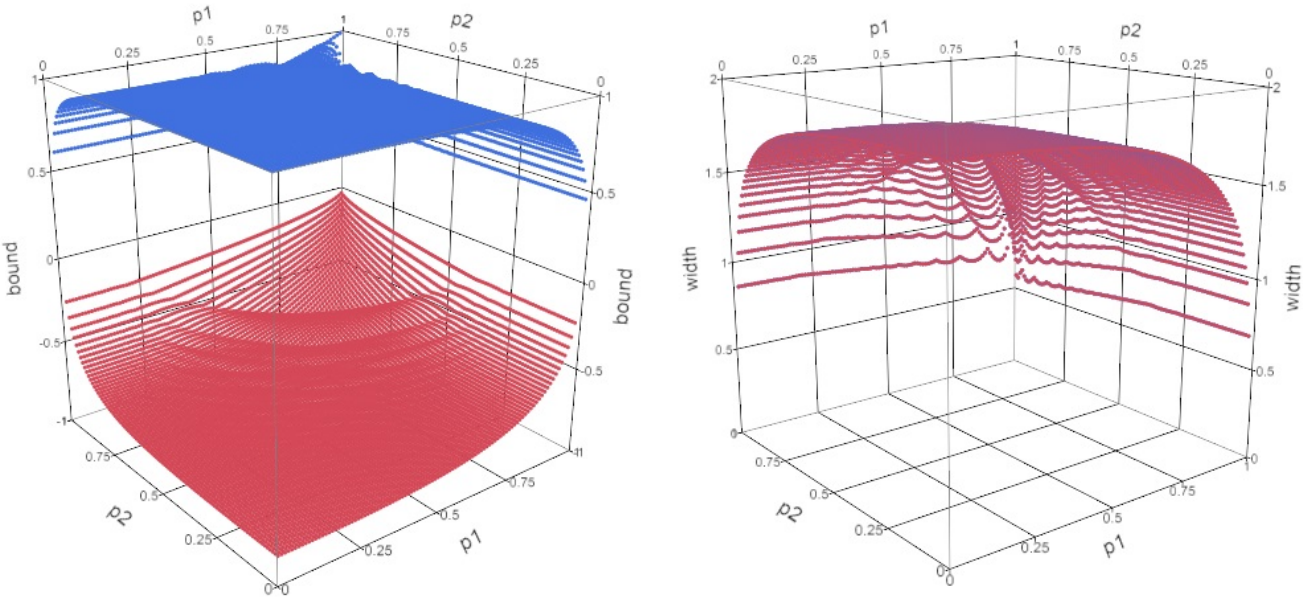


(b) Interval Widths

Fig. 4.23. Case: $r_1 = 2$, $r_2 = 2$

[1.379, 1.735], respectively.

Case: $r_1 = 2, r_2 = 4$



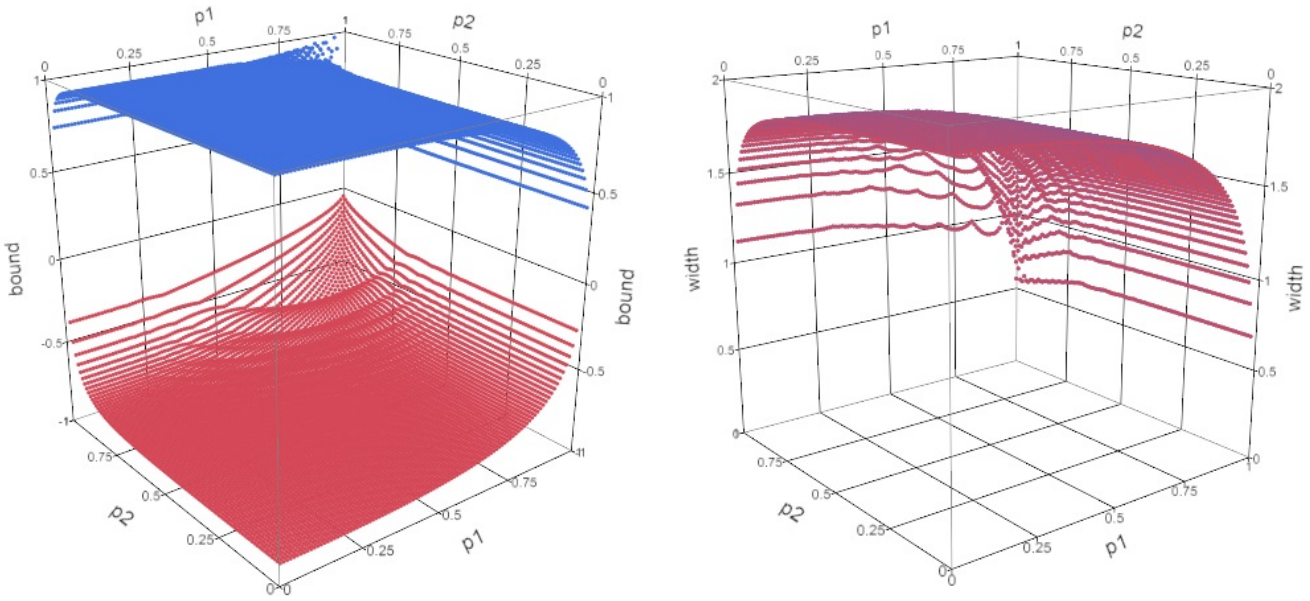
(a) Fréchet Bounds (b) Interval Widths

Fig. 4.24. Case: $r_1 = 2, r_2 = 4$

Here, the shape shifts away from symmetry again, with the dorsal fin and ridges near the $p_1 = p_2$ plane, but not lying along it. The dorsal fin and ridges also appear shallower. The upper bound had a maximum of 1.00 (0.996) at (0.02, 0.03) and the lower bound had a minimum of -0.85 at (0.01, 0.01). The width had a maximum of 1.85 and a minimum of 0.58, with median and IQR of 1.71 and [1.500, 1.802], respectively.

Case: $r_1 = 2, r_2 = 10$

As in the previous case, the shape of the bounds continues to shift away from center and the dorsal fin and ridges become even shallower. The upper bound had a maximum of 0.98 at (0.02, 0.04) and the lower bound had a minimum of -0.89 at (0.01, 0.01). The width had a maximum of 1.88 and a minimum of 0.57, with median and IQR of 1.78 and [1.595, 1.848], respectively.



(a) Fréchet Bounds (b) Interval Widths

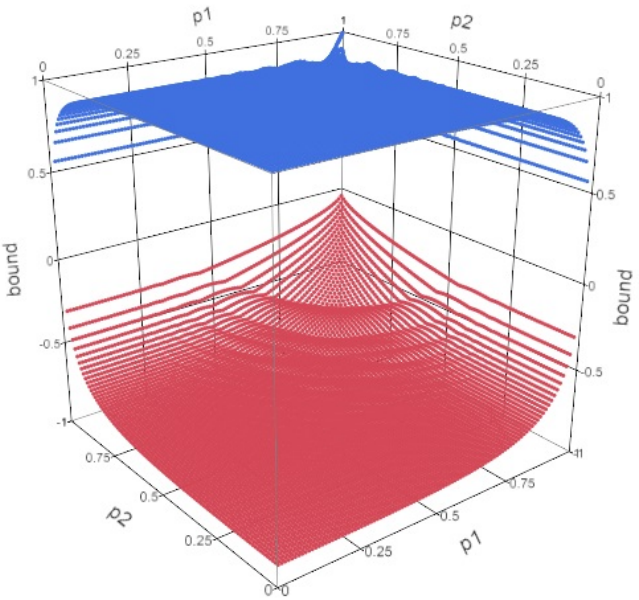
Fig. 4.25. Case: $r_1 = 2$, $r_2 = 10$

Case: $r_1 = 4$, $r_2 = 4$

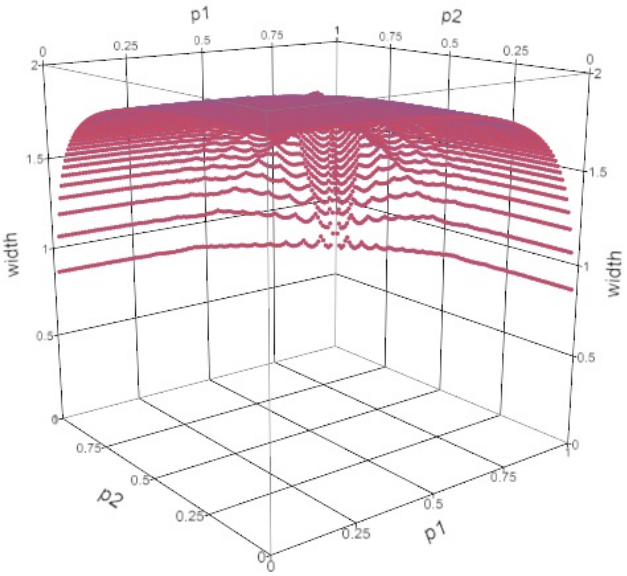
Returning to symmetry about the plane $p_1 = p_2$, in this case, the bounds have larger intervals and the ridges and dorsal fin not nearly as pronounced as in the previous symmetric cases. The upper bound had a maximum of 1 at (0.99, 0.99) and the lower bound had a minimum of -0.89 at (0.01, 0.01). The width had a maximum of 1.89 and a minimum of 0.70, with median and IQR of 1.80 and [1.627, 1.867], respectively.

Case: $r_1 = 4$, $r_2 = 10$

It appears that as r_1 and r_2 increase, the bounds become wider and the ridges become shallower, and in the lower bounds, the ridges tend to move toward the corner where p_1 and p_2 approach 1. The bounds are not symmetrical here, shifted off of the plane $p_1 = p_2$. The upper bound had a maximum of 1.00 (0.996) at (0.03, 0.04) and the lower bound had a minimum of -0.93 at (0.01, 0.02). The width had a maximum of 1.92 and a minimum of 0.69, with median and IQR of

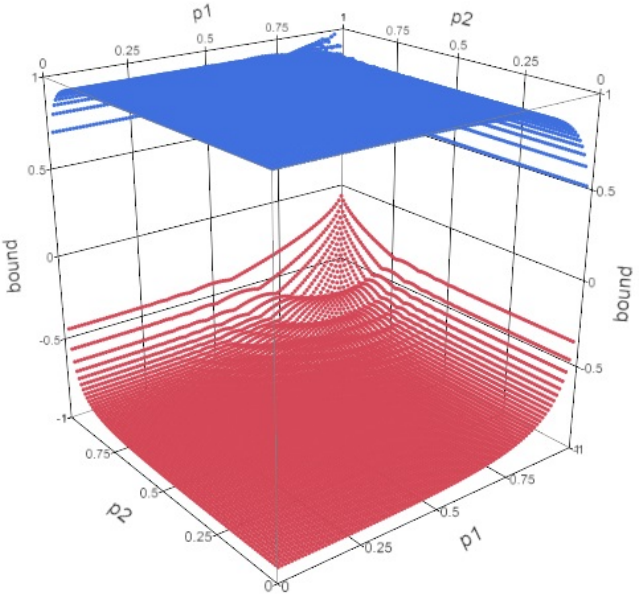


(a) Fréchet Bounds

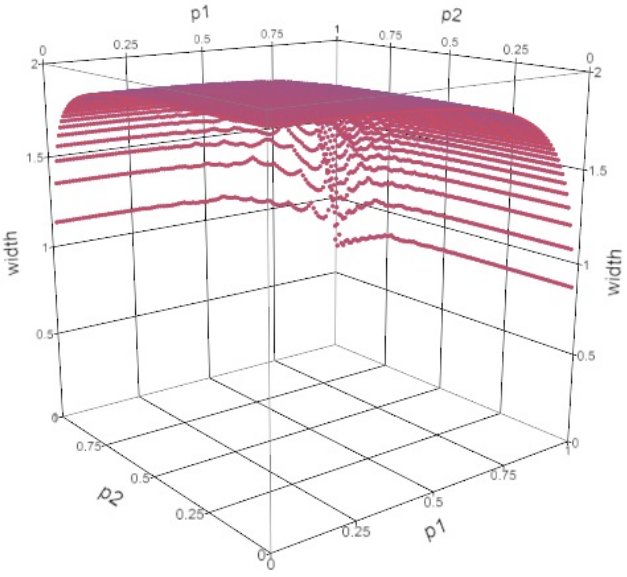


(b) Interval Widths

Fig. 4.26. Case: $r_1 = 4$, $r_2 = 4$



(a) Fréchet Bounds

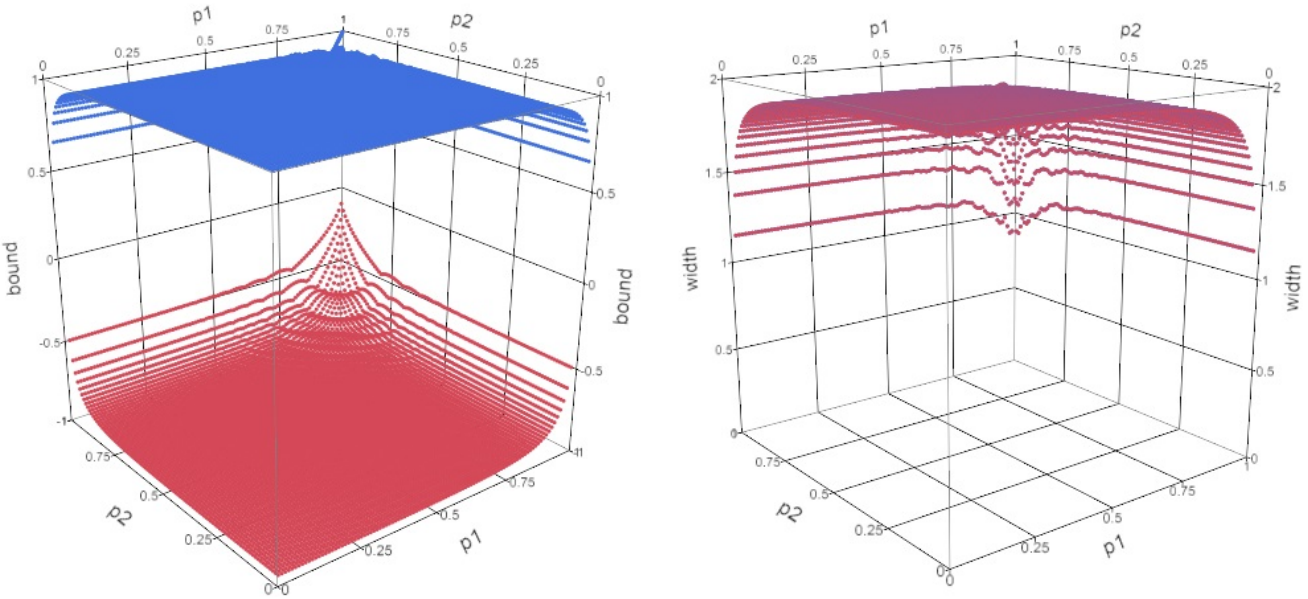


(b) Interval Widths

Fig. 4.27. Case: $r_1 = 4$, $r_2 = 10$

1.86 and [1.739, 1.925], respectively.

Case: $r_1 = 10$, $r_2 = 10$



(a) Fréchet Bounds (b) Interval Widths

Fig. 4.28. Case: $r_1 = 10$, $r_2 = 10$

Again, the bounds have symmetry since $r_1 = r_2$. The upper bound has a maximum of 1 at (0.99, 0.99) and the lower bound had a minimum of -0.96 at (0.01, 0.01). The width has a maximum of 1.96 and a minimum of 0.88, with median and IQR of 1.92 and [1.850, 1.947], respectively.

Remarks Regarding Two-Variable Cases

As r_1 and r_2 increase, the widths of the bounds also increase. The upper bounds reach the maximum of 1 when $r_1 = r_2$, and the minimum lower bound is often at the corner (0.01, 0.01). The graphs generally have an "open mouth" shape with the larger intervals near the corner (0.01, 0.01), with a gradual shrinking of the intervals as p_1 and p_2 increase, then shrinking rapidly as p_1 and p_2 approach the corner (0.99, 0.99). This shape is due to the definition of the p_i as the probability of failure and the nature of the negative binomial distribution. An exception to this is the points along

the plane $p_1 = p_2$, where in many cases there is a "keel" or "dorsal fin" on the upper bound, where it is 1 or nearly 1.

4.3.2 Three-Variable Cases

As examples of Fréchet bounds in three-variable cases, certain combinations of 2 and 10 were chosen for r_1, r_2, and r_3. The bounds are shown for all sets of p_1 and p_2, starting at 0.01 and ending at 0.99 in increments of 0.01 for both probabilities, while p_3 was restricted to 0.25, 0.50, and 0.75.

In the AR(1) case, the Fréchet bounds are slightly trickier since $\rho_{12} = \rho_{23} = \rho$ and $\rho_{13} = \rho^2$. Often in the AR(1) case, only positive correlations are considered due to the exponent on ρ_{13}, but the possibility of negative correlations will be allowed for the sake of completeness. In the above equation (4.1), the quantities $\rho^*_{13L} = S(\rho_{13L})\sqrt{\rho_{13L}}$ and $\rho^*_{13U} = S(\rho_{13U})\sqrt{\rho_{13U}}$, where $S(\cdot)$ is the signum function.

Case (CS): $r_1 = 2, r_2 = 2, r_3 = 2$

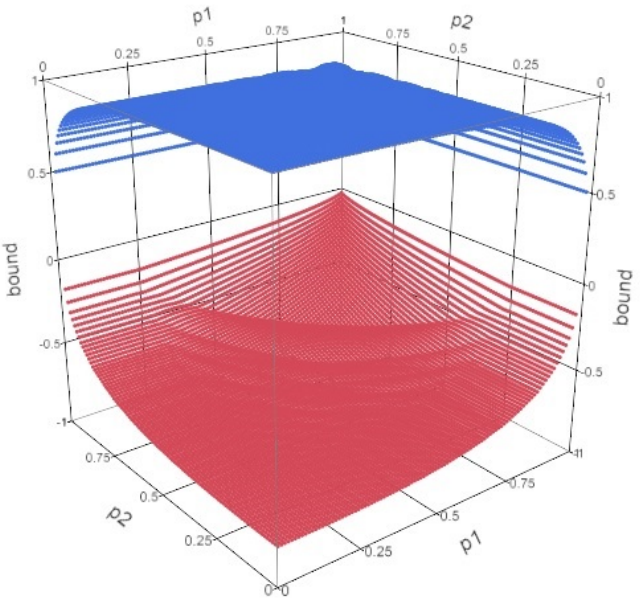
Though it may be difficult to see in Figure 4.29, the ridges in the lower bounds shift toward (p_1, p_2)=(0, 0) and get further apart as p_3 increases. The upper and lower bounds are symmetric about the plane $p_1 = p_2$ for each p_3. The dorsal fin appears as seen in the two-variable cases, but it is not nearly as prominent except in Figure 4.29c.

Case (CS): $r_1 = 2, r_2 = 2, r_3 = 10$

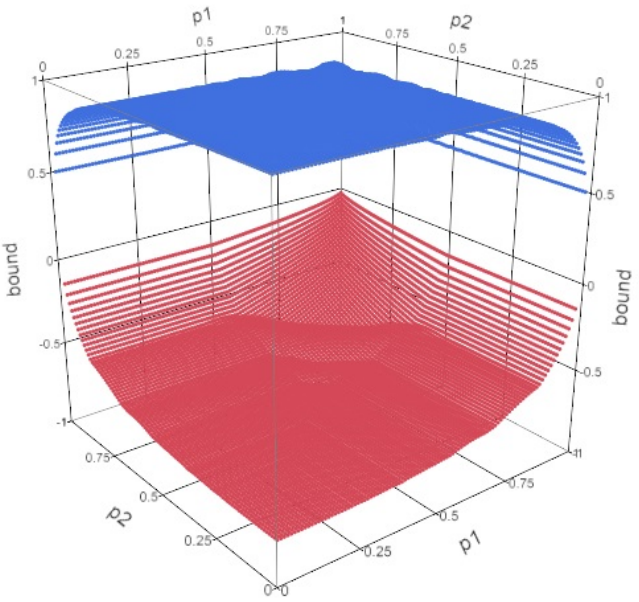
The upper bounds in this case appear nearly completely smooth, except for a slight ripple in Figure 4.30c, indicating shallow ridges. The lower bounds appear to be nearly the same in all three figures.

Case (CS): $r_1 = 2, r_2 = 10, r_3 = 10$

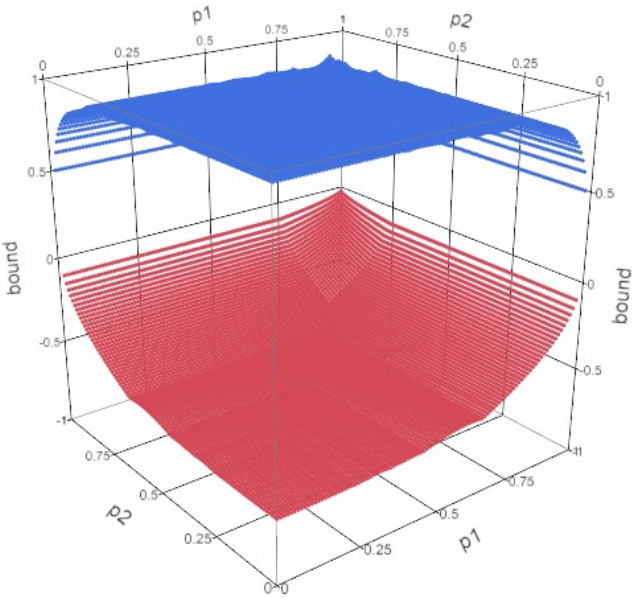
In this case, the bounds take on a shifted shape, as though the figures in the previous case had been twisted. The upper bounds in Figures 4.31a and 4.31b again appear smooth, and in Figure



(a) $p_3 = 0.25$

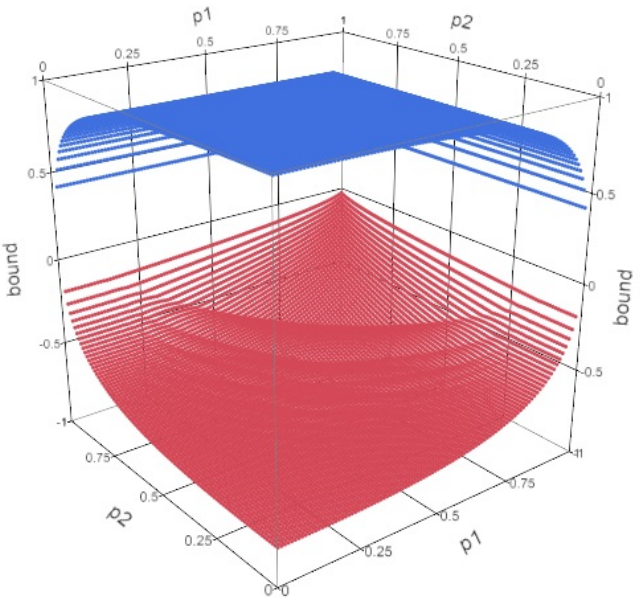


(b) $p_3 = 0.50$

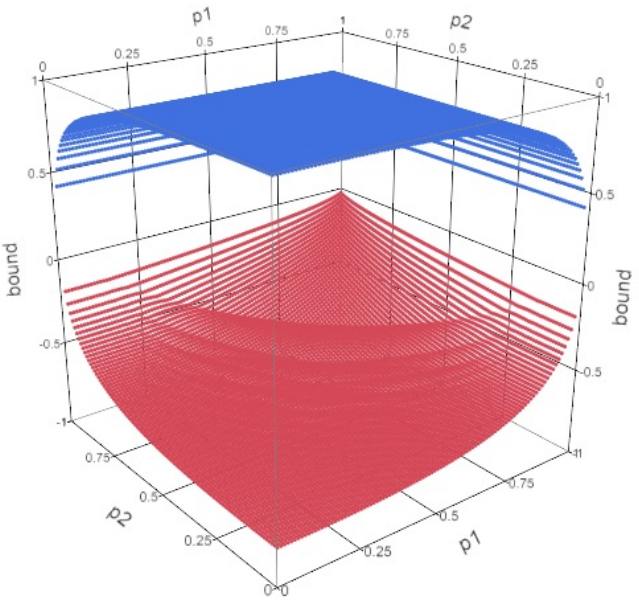


(c) $p_3 = 0.75$

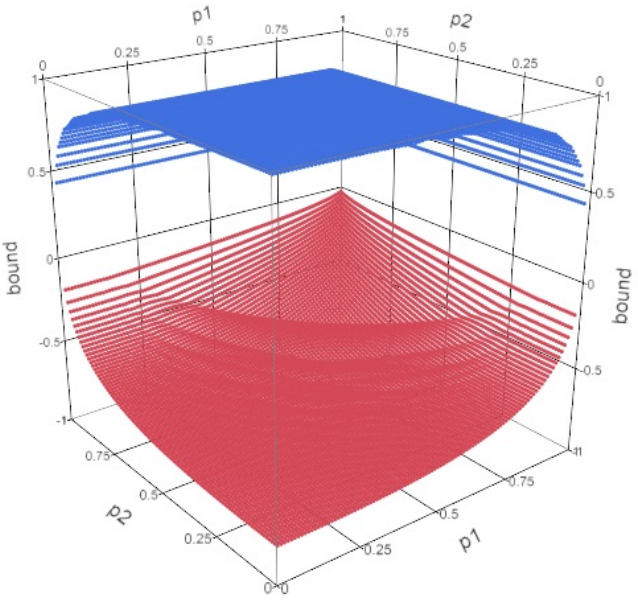
Fig. 4.29. Case (CS): $r_1 = 2$, $r_2 = 2$, $r_3 = 2$



(a) $p_3 = 0.25$

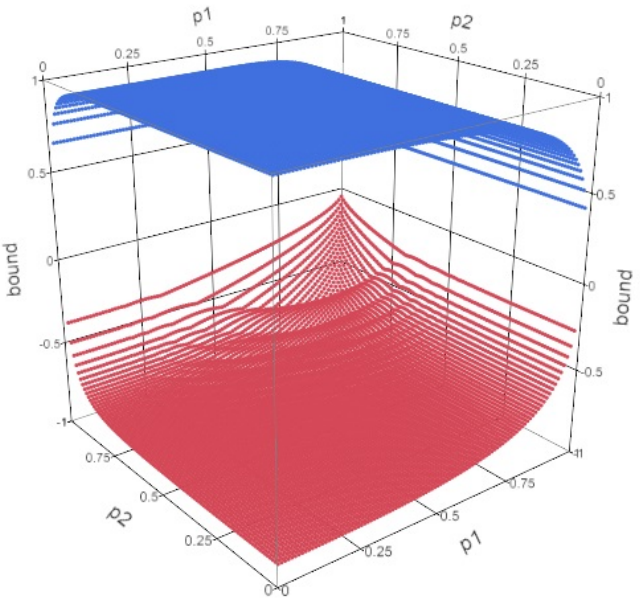


(b) $p_3 = 0.50$

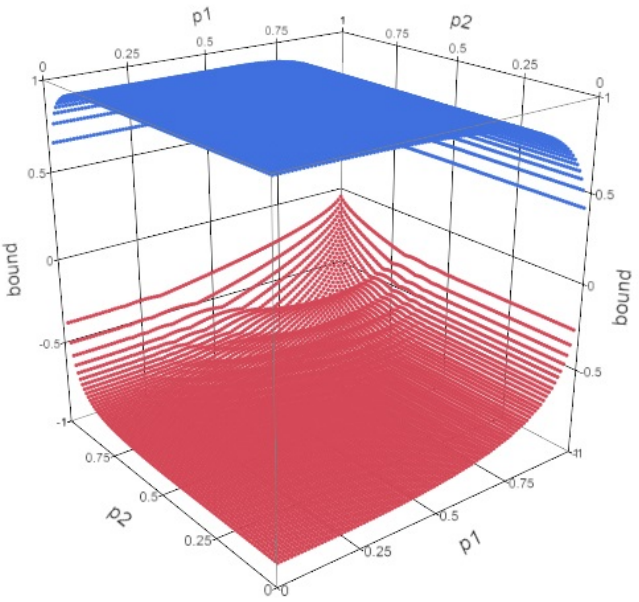


(c) $p_3 = 0.75$

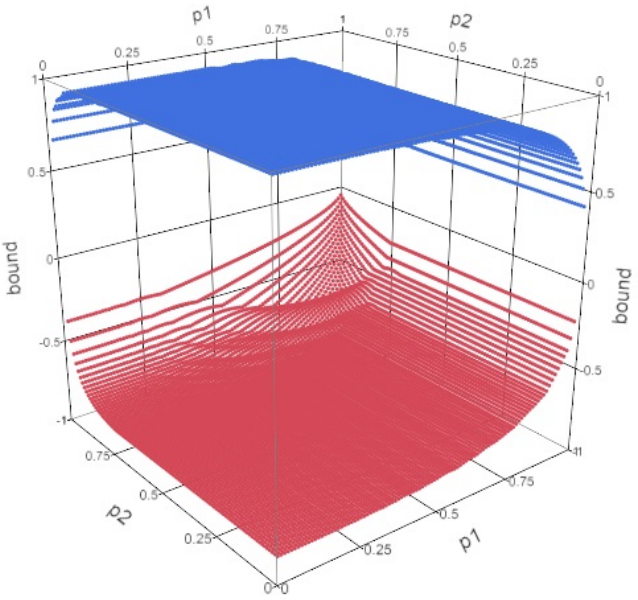
Fig. 4.30. Case (CS): $r_1 = 2$, $r_2 = 2$, $r_3 = 10$



(a) $p_3 = 0.25$



(b) $p_3 = 0.50$



(c) $p_3 = 0.75$

Fig. 4.31. Case (CS): $r_1 = 2$, $r_2 = 10$, $r_3 = 10$

4.31c, there appears to be a few shallow ridges.

Case (CS): $r_1 = 10$, $r_2 = 10$, $r_3 = 10$

The symmetry returns in this case, and so do the ridges in the upper bounds, the most prominent where p_3=0.75, though not by much. The lower bounds still appear similar.

Case (AR(1)): $r_1 = 2$, $r_2 = 2$, $r_3 = 2$

The AR(1) cases appear similar to the CS cases, but they appear as if someone took the CS cases and played with the ridges, smoothing them in different places and shifting them about. The ridges in the upper bounds now have disjoints, and are no longer symmetric where $r_1 = r_2$. In the lower bounds for this case, the ridges seen in the CS case now fail to appear where $p_1 \leq p_3$, but the ridges seem the same as the CS case where $p_1 > p_3$.

Case (AR(1)): $r_1 = 2$, $r_2 = 2$, $r_3 = 10$

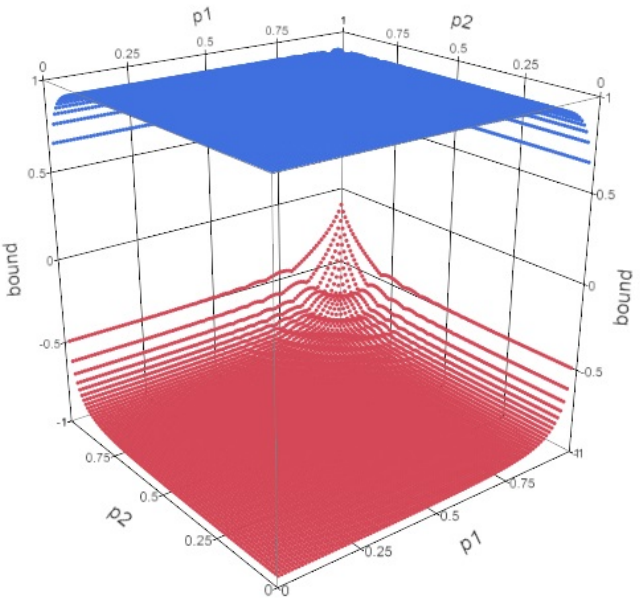
In this case, the lower bounds appear the same as in 4.30, but the upper bounds have ridges, though they are shifted off to the right side of the figures. The ridges appear similar in Figures 4.34a and 4.34b, but they are slightly sharper in Figure 4.34c.

Case (AR(1)): $r_1 = 2$, $r_2 = 10$, $r_3 = 10$

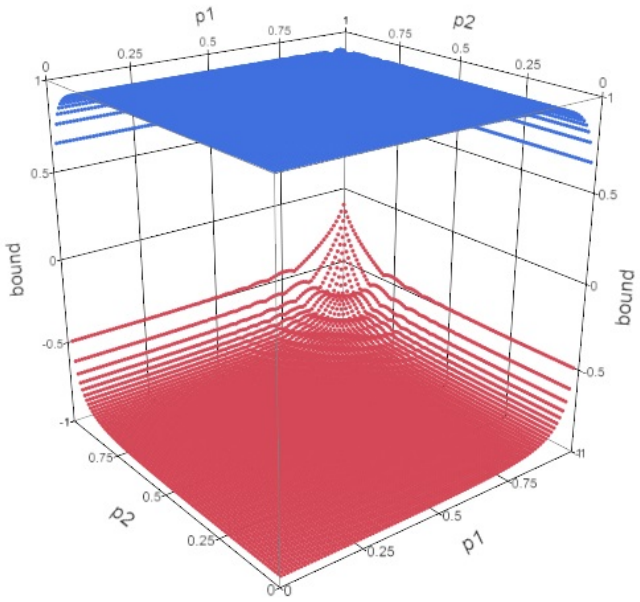
In this case, the upper and lower bounds differ little between the different values of p_3. The figures also appear similar to Figure 4.31, except where p_3=0.75. In the CS case, the lower bounds are not as smooth as p_1 increases.

Case (AR(1)): $r_1 = 10$, $r_2 = 10$, $r_3 = 10$

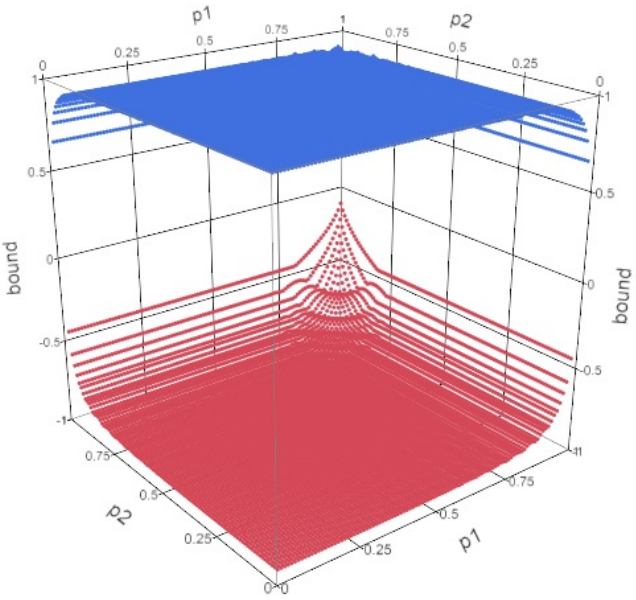
The bounds in this case appear nearly the same as in the CS case, with some slight differences in the ridges of the upper bounds.



(a) $p_3 = 0.25$

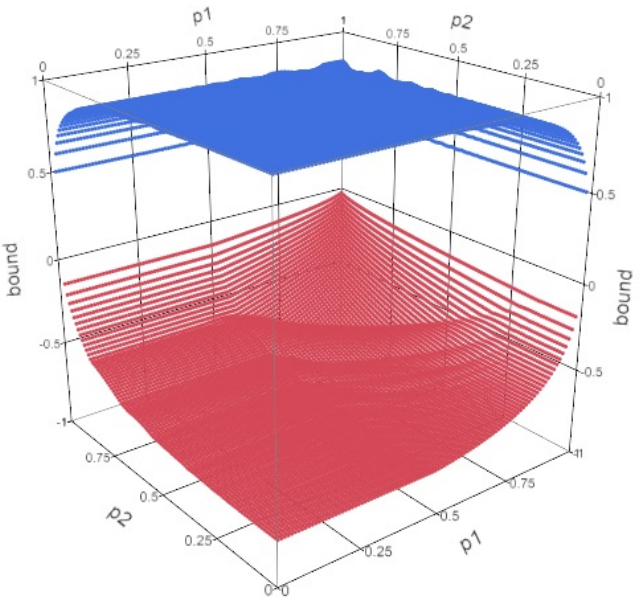


(b) $p_3 = 0.50$

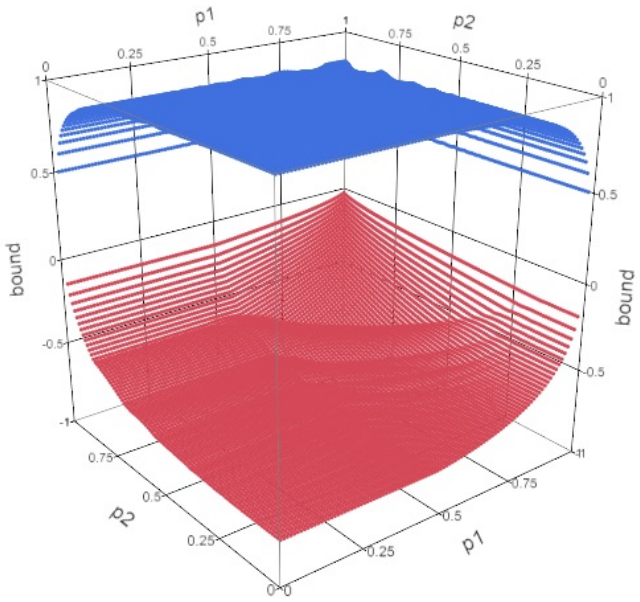


(c) $p_3 = 0.75$

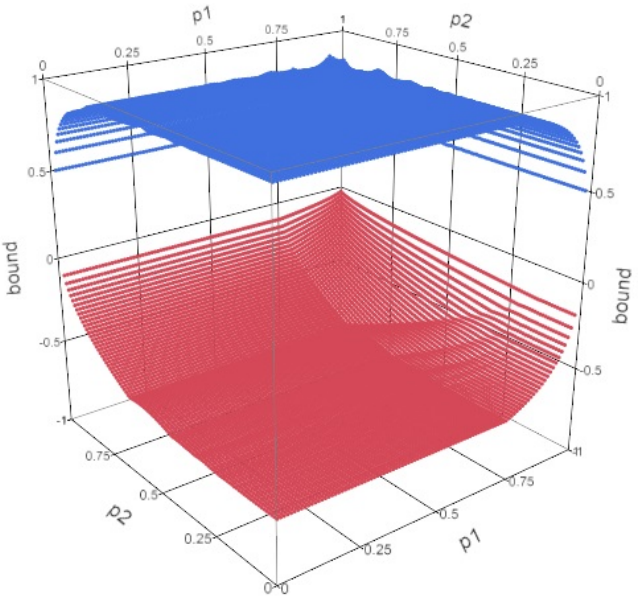
Fig. 4.32. Case (CS): $r_1 = 10$, $r_2 = 10$, $r_3 = 10$



(a) $p_3 = 0.25$

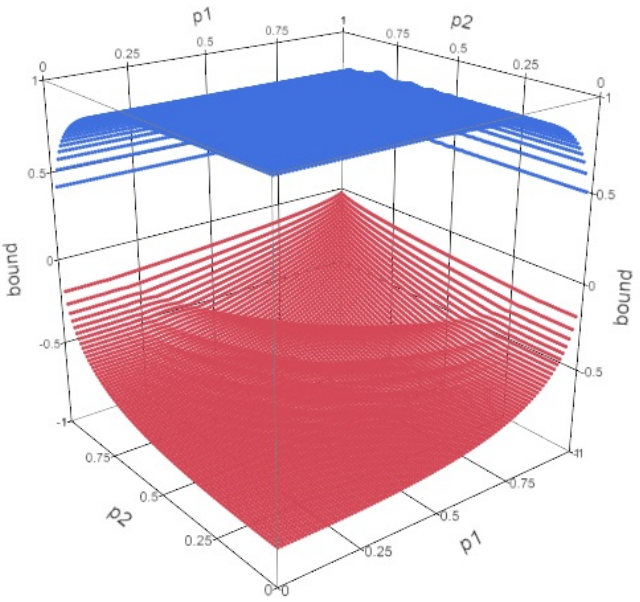


(b) $p_3 = 0.50$

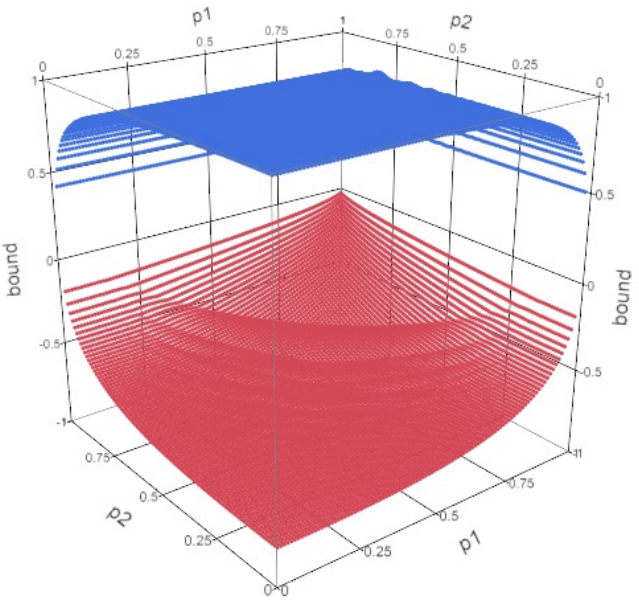


(c) $p_3 = 0.75$

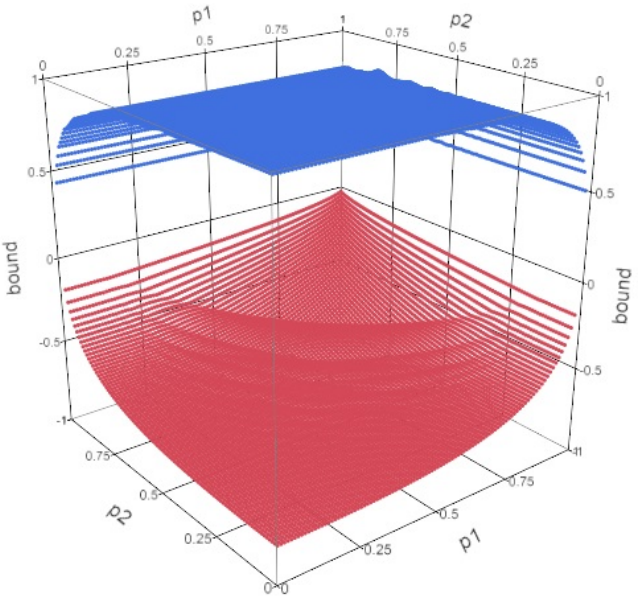
Fig. 4.33. Case (AR(1)): $r_1 = 2$, $r_2 = 2$, $r_3 = 2$



(a) $p_3 = 0.25$

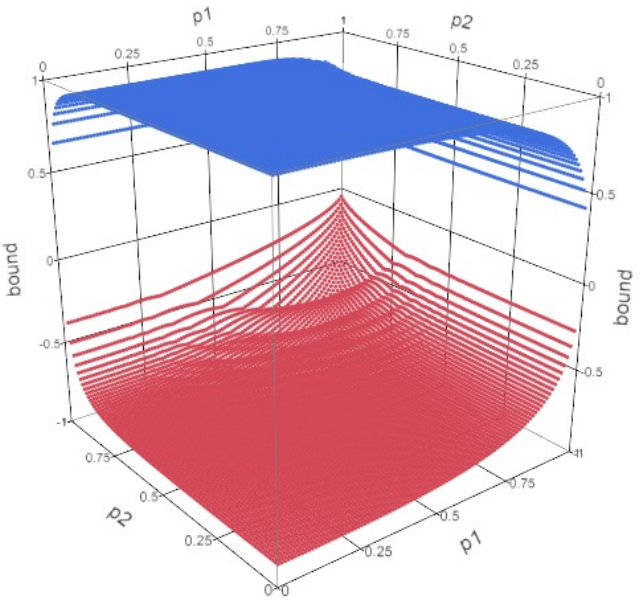


(b) $p_3 = 0.50$

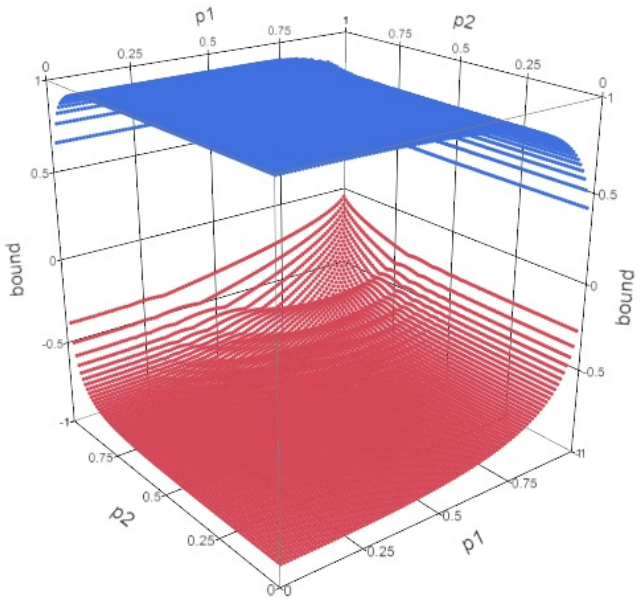


(c) $p_3 = 0.75$

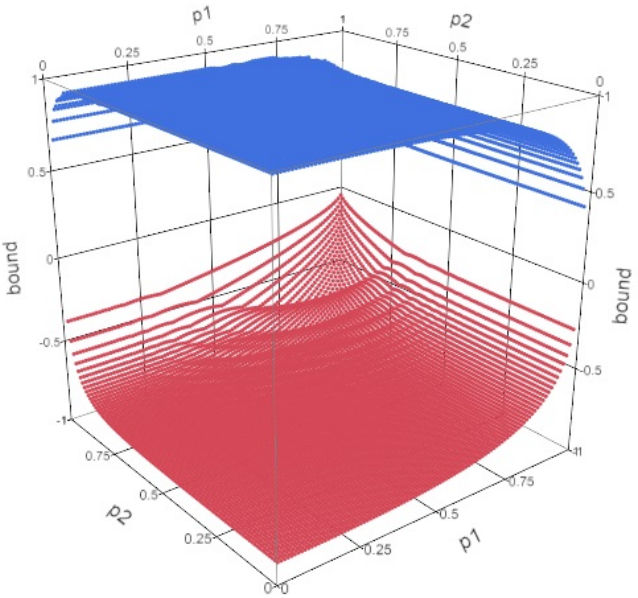
Fig. 4.34. Case (AR(1)): $r_1 = 2$, $r_2 = 2$, $r_3 = 10$



(a) $p_3 = 0.25$

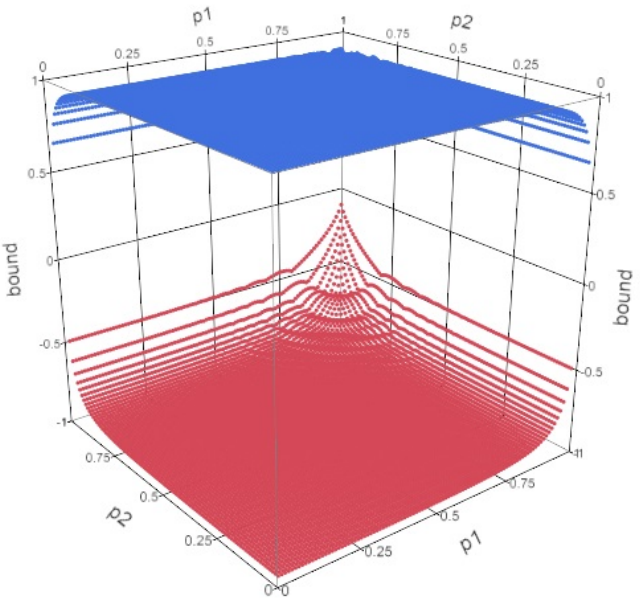


(b) $p_3 = 0.50$

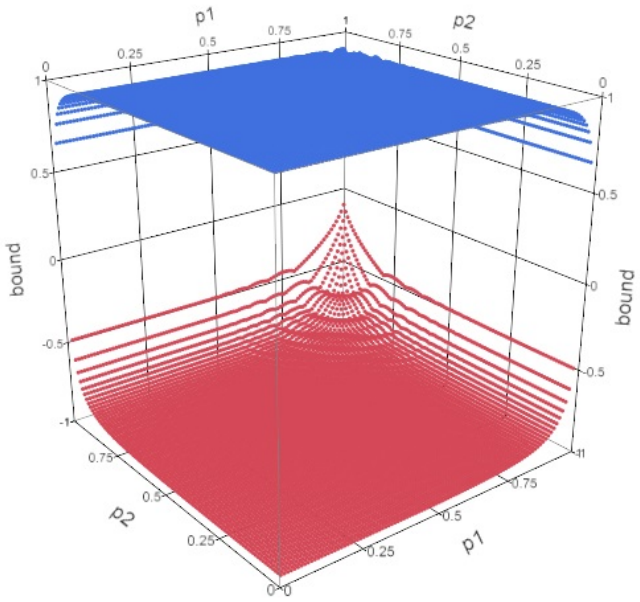


(c) $p_3 = 0.75$

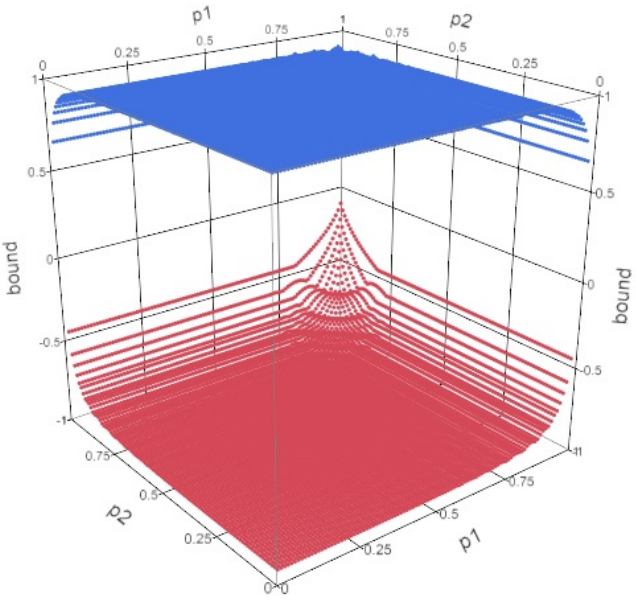
Fig. 4.35. Case (AR(1)): $r_1 = 2$, $r_2 = 10$, $r_3 = 10$



(a) $p_3 = 0.25$



(b) $p_3 = 0.50$



(c) $p_3 = 0.75$

Fig. 4.36. Case (AR(1)): $r_1 = 10$, $r_2 = 10$, $r_3 = 10$

Remarks Regarding Three-Variable Cases

For both CS and AR(1) cases, the widest intervals appear near the corner where p_1 and p_2 are near or at zero. This is likely due to the definitions of the marginal probabilities as the probabilities of failure rather than the probabilities of success. The ridges in the upper and lower bounds are due to the discrete nature of the negative binomial distribution. These cases appear to shift less than the similar binomial cases as p_3 shifts (compare Figures 4.29 - 4.32 to Figures 4.11 - 4.14 for CS and compare Figures 4.33 - 4.36) to Figures 4.15 - 4.18)). Patterns can be seen in the CS cases in that symmetry occurs when $r_1 = r_2 = r_3$, and as the r_i increase, so do the widths of the intervals. Asymmetry is most obvious where $r_1 \neq r_2$, and the figure is shifted toward the side where $p_2 = 0.99$. In the AR(1) cases, asymmetry is most obvious where $r_1 \neq r_2$, however, none of the cases are symmetric because of the extra limitation the AR(1) structure requires. As the r_i increase, symmetry appears to return, but the figures are never quite symmetric.

4.4 NHANES Analysis

As an example of how the Fréchet bounds should be used in a real-world situation, data from the National Health and Nutrition Survey (NHANES) [3] [4] [5] were analyzed. The data for environmental phenols (Y_1) and environmental pesticides (Y_2) from 2009-2010 were used in a simple GEE analysis of whether males and females differed in the number of environmental phenols and pesticides with a "high" measurement. The environmental phenols (Y_1) included eight analytes: 4-tert-octylphenol (ng/mL), benzophenone-3 (ng/mL), bisphenol A (ng/mL), triclosan (ng/mL), butyl paraben (ng/ml), ethyl paraben (ng/ml), methyl paraben (ng/ml), and propyl paraben (ng/ml). The environmental pesticides (Y_2) included five analytes: 2,5-dichlorophenol (μg/L), O-Phenyl phenol (μg/L), 2,4-dichlorophenol (μg/L), 2,4,5-trichlorophenol (μg/L), and 2,4,6-trichlorophenol (μg/L).

4.4.1 Methods

The data were retrieved from the NHANES website. All measurements on the analytes used were derived from urine specimens. The results of the analytes were divided into two groups,

either "high" (>75th percentile) or "not high" (≤75th percentile), which were coded as 1 and 0, respectively. The calculation of the 75th percentile included the values substituted for results that were below the lowest level of detection. Subjects with any missing analytes from either the environmental phenols or the environmental pesticides were excluded from the analysis. These binary outcomes were summed according to whether they were from the environmental phenol or environmental pesticide datasets, creating two binomial outcomes for each subject. The Fréchet bounds were calculated using PROC FREQ and PROC IML in order to find the sample marginal probabilities, ρ_{12L}, and ρ_{12U} as described in Section 4.2. The binomial variables by the subject identifier were submitted to PROC GENMOD with gender and type of analyte (phenol or pesticide) along with the interaction between the two (gender*type) as predictor variables using a logit link for the generalized linear model under a reference cell parameterization Female and Phenols are the reference levels of their respective predictor variables. The working correlation from the resulting GEE analysis was recorded along with parameter estimates and inference on the predictor variables. The working correlation from the analysis was compared to the calculated Fréchet bounds.

4.4.2 Results

There were 2819 subjects each present in the environmental phenols and environmental pesticides datasets. Out of these, 2749 (97.5%) subjects had all results for all analytes and were included in the analysis. There were 1399 (50.9%) males and 1350 (49.1%) females. The right-tail cumulative probabilities were

$$P(y_1 \geq k) = \sum_{s=k}^{8} \binom{8}{s} \hat{p}_1^s \hat{q}_1^{8-s}$$

$$P(y_2 \geq l) = \sum_{t=l}^{5} \binom{5}{t} \hat{p}_2^t \hat{q}_2^{5-t}$$

The estimated marginal probability for the environmental phenols was $\hat{p}_1 = 0.232$, and for the environmental pesticides it was $\hat{p}_2 = 0.222$. Thus, the Fréchet bounds were calculated as

$$\hat{\rho}_{12L} = \frac{\sum_{k=1}^{8}\sum_{l=1}^{5}\max\left[\sum_{s=k}^{8}\binom{8}{s}(0.232)^s(0.768)^{8-s}+\sum_{t=l}^{5}\binom{5}{t}(0.222)^t(0.778)^{5-t}-1,\,0\right]-8(0.232)5(0.222)}{(8(0.232)(0.768)5(0.222)(0.778))^{1/2}}$$

$$\hat{\rho}_{12U} = \frac{\sum_{k=1}^{8}\sum_{l=1}^{5}\min\left[\sum_{s=k}^{8}\binom{8}{s}(0.232)^s(0.768)^{8-s},\,\sum_{t=l}^{5}\binom{5}{t}(0.222)^t(0.778)^{5-t}\right]-8(0.232)5(0.222)}{(8(0.232)(0.768)5(0.222)(0.778))^{1/2}}$$

which resulted in $[\hat{\rho}_{12L}, \hat{\rho}_{12U}]$ = [-0.8812, 0.9222].

The GEE procedure in PROC GENMOD produced a working correlation of $\hat{\rho}_{12}$ = 0.2618, which was well within the Fréchet bounds calculated. Since $\hat{\rho}_{12}$ was within the Fréchet bounds, the inference on whether there is a difference between males and females is assumed to be valid. The p-value for the gender*type parameter is statistically significant (χ^2=117.48, $p < 0.0001$), meaning that there is a statistical difference between at least two of the four groups (Male & Pesticides, Male & Phenols, Female & Pesticides, and Female & Phenols) after the assumptions of the model have been met. Because this interaction term is significant, the potential difference between males and females cannot be considered separately from the type of analyte being examined. The parameter estimates, empirical standard errors, associated Wald confidence intervals, Type 3 score test statistic, and p-values are seen in Table 4.1.

Remarks Regarding NHANES Results

This analysis is particularly simple and limited in that the result is not adjusted for any factors that may be influential in the rates of exposure to these chemicals (e.g. age, race, socio-economic status, career, etc.). If the working correlation had fallen outside of the boundaries, the analysis would have needed to be adjusted by either specifying the working correlation so that it would be within the bounds or by choosing a different analysis method, such as the multivariate probit as suggested by Sabo and Chaganty [10].

Table 4.1. NHANES Parameter estimates, Empirical Standard Errors, Wald Confidence Intervals, and Inference

Parameter	Estimate	Empirical Standard Error	95% Wald Confidence Interval	Type 3 Score Test	p-value
Intercept	-0.882	0.0302	[-0.9412, -0.8226]	–	–
Gender	-0.680	0.0444	[-0.7673, -0.5932]	72.18	<0.0001
Type	-0.369	0.0426	[-0.4526, -0.2856]	1.07	0.3018
Gender*Type	0.676	0.0607	[0.5566, 0.7947]	117.48	<0.0001

CHAPTER 5

DISCUSSION

The Fréchet bounds are an important feature of simulations and analyses involving discrete dependent variables, as demonstrated in the preceding chapters.

5.1 Summary of Findings

Chapter 2 demonstrated the difference between two methods of simulating binary data with specified marginal probabilities and correlation while staying within the Fréchet bounds. While generally similar concerning the parameters of interest, the multinomial sampling (MS) technique came out ahead of the technique by Emrich and Piedmonte (EP) [6] in a few key areas. The EP method is limited by the requirement of the bivariate normal distribution of an invertible covariance (and hence correlation) matrix. The MS method does not require this, so all combinations of the marginal probabilities and the target correlations within the Fréchet bounds may be used as simulation parameters. In the simulated pre-/post-treatment analysis with two repeated measures, the MS method of simulation more often resulted in the correct inference on the research question. The simplicity of the MS method over the EP method and the direct link of the binary joint cdf to the results of the simulations would also be advantages of the MS method. Also, the EP method of simulation took more CPU time than the MS method, though that may possibly be due to inefficient programming on the part of the author.

Chapter 3 demonstrated that the multinomial sampling method for simulating binary dependent datasets could be utilized with a specified odds ratio instead of a specified correlation. This cannot be done using the EP method, as the method depends upon the specification of a correlation. Many different datasets can result in the same odds ratio; it is not as limiting as the correlation. The Fréchet bounds on the odds ratio for both the case of common odds ratio (i.e. $\psi_{12} = \psi_{13} = \psi_{23} = \psi$) and the case of unstructured odds ratios (ψ_{12} and ψ_{23} allowed a range of $[0, \infty)$ with ψ_{13} limited by

more stringent bounds) were described, and datasets were simulated which incorporated the odds ratios within the bounds. As would be expected, the common odds ratio estimates followed similar changes in the measures of interest across all estimates of the odds ratio. As might be expected, the measures of interest for the unstructured case stayed the same for the estimates of ψ_{12} and ψ_{23} while the same measures varied for the estimate of ψ_{13}.

Chapter 4 explored the Fréchet bounds for the binomial and negative binomial distributions in two-variable cases and three-variable compound symmetric and first-order auto-regressive cases. The overall forms of the families of Fréchet bounds were described. A general pattern was seen; as the n_i (binomial) or r_i (negative binomial) increased, the Fréchet bounds became wider. This makes intuitive sense due to the Central Limit Theorem. Ridges and "bumps" in the appearance of the bounds were attributed to the discreteness of the distributions, especially since the ridges seen in the binomial cases matched the n_i being used. The example involving NHANES environmental phenols and environmental pesticides showed a statistically significant difference between males and females. The analysis did not have to be adjusted because the working correlation produced by the GEE procedure was within the calculated Fréchet bounds.

5.2 Limitations

This work is limited in scope, having only provided a few examples of the possible families of distributions affected by the Fréchet bounds, which was done for efficiency of presentation. There are other methods of simulating binary data which were not compared to the multinomial sampling method. However, since the EP method is the most popular method of simulating binary data, it was deemed suitable to compare the MS method only to this one. The simulated two-group pre-/post-treatment study in Chapter 2 and the NHANES analysis in Chapter 4 are simplistic though realistic.

5.3 Immediate Extensions

Immediate extensions of this work would include:

- comparison of the multinomial sampling technique to simulation techniques other than the EP method,
- development of a simulation technique suitable for other discrete distributions,
- calculation and examination of the odds ratio bounds for discrete distributions other than the binary distribution,
- examination of the Fréchet bounds of more families of binomial and negative binomial distributions, and
- demonstrations of difficulties when ignoring the bounds similar to the analysis in the 2010 paper by Sabo and Chaganty [10] to greater solidify the importance of the Fréchet bounds.

Appendix A

SAS CODE RELEVANT TO CHAPTER 2

A.1 Two-variable dependent binary Emrich and Piedmonte [6] technique as described in Section 2.3.1

```
proc iml;
/* Create function normcorr to find the correlation for a standard
bivariate normal that corresponds to given bivariate binary
correlation and marginal probabilities */
start normcorr(k,pa,pb);
/* Right-hand side of eq.  2.1 in Emrich and Piedmonte paper */
rhs=k*(pa*(1-pa)*pb*(1-pb))**(1/2)+pa*pb;
/* Calculate the difference between rhs and the cdf using all
possible correlations */
do c=-0.999 to .999 by .001;
normcdf=probbnrm(quantile('NORMAL',pa),quantile('NORMAL',pb),c);
diff = abs(normcdf-rhs);
difvec=difvec//diff;
cvec=cvec//c;
end;
/* Choose the correlation that corresponds to the difference closest
to zero */
mindif=difvec[>:<];
corr=cvec[mindif];
return(corr);
```

```
finish normcorr;
/* Set up marginal probabilities (p1 and p2), q1 and q2 (q1=1-p1,
q2=1-p2), target correlation (k), and number of observations to
determine a mean rho (n) - before the next step */
/* Frechet bounds */
Lp1p2=max(-sqrt((p1*p2)/(q1*q2)),-sqrt((q1*q2)/(p1*p2)));
Up1p2=min(sqrt((p1*q2)/(q1*p2)),sqrt((q1*p2)/(p1*q2)));
/* Find correlation for standard bivariate normal that corresponds
to k for bivariate binary distribution */
corr=normcorr(k,p1,p2);
/* Begin simulations */
m=10000;
cat=j(m,1,0);  /* Column for categorizing whether the estimated
correlation falls within the Frechet bounds */
rho=j(m,1,0); /* Column for the estimated correlation */
ckv=j(m,1,0); /* Column for marking whether an observed variance is
zero */
do j=1 to m;
/* Create multivariate random observations */
mean=0,0;
var=1,1;
varcov = ((1 || corr)//(corr || 1));
y1=j(n,1);
y2=j(n,1);
do l=1 to n;
call randseed(47);
sim=randnormal(n,mean,varcov);
```

```
if sim[l,1] <= quantile('NORMAL',p1) then y1[l]=1;
else y1[l]=0;
if sim[l,2] <= quantile('NORMAL',p2) then y2[l]=1;
else y2[l]=0;
end;
```

A.2 Two-variable dependent binary multinomial sampling technique as described in Section 2.3.2

```
proc iml;
/* Set up marginal probabilities (p1 and p2), q1 and q2 (q1=1-p1,
q2=1-p2), target correlation (a), and number of observations to
determine a mean rho (n) - before the next step */
/* Frechet bounds */
Lp1p2=max(-sqrt((p1*p2)/(q1*q2)),-sqrt((q1*q2)/(p1*p2)));
Up1p2=min(sqrt((p1*q2)/(q1*p2)),sqrt((q1*p2)/(p1*q2)));
/* Create a correlation matrix with a as the target correlation
between the two RVs */
corr=(1 || a) // (a || 1);
/* Calculate E[XY], that is, the probability that both variables are
successes using the formula for calculating the Pearson correlation,
solved for E[XY] instead of CORR */
p12=p1*p2+corr[1,2]*sqrt(p1*q1)*sqrt(p2*q2);
/* Set up the probability mass function */
p11=p12; /* prob.  of two successes */
p10=p1-p12; /* prob.  of 1st success and 2nd failing */
p01=p2-p12; /* prob.  of 1st failing and 2nd success */
p00=1-p1-p2+p12; /* prob.  of two failures */
```

```
/* Create the CDF of the distribution */
/* pa, pb, and pc are for calculating which 'slot' a Uniform(0,1)
outcome will go in */
pa=p11;
pb=p11+p10;
pc=p11+p10+p01;
m=10000;
/* creating column vectors of zeroes with m rows */
cat=j(m,1,0);  /* Column for categorizing whether the estimated
correlation falls within the Frechet bounds */
rho=j(m,1,0); /* Column for the estimated correlation */
ckv=j(m,1,0); /* Column for marking whether an observed variance is
zero */
/* Begin DO loop for simulations */
do j=1 to m;
vec=j(n,2,0); /* creating nx2 matrix of zeroes */
do i=1 to n;
seed=47; /* random seed for calculating Uniform(0,1) random variable
*/
c=j(1,1,seed); /* 1x1 matrix seed */
/* Assigning bivariate binary outcomes to uniform random variable */
u=uniform(c);
if u<=pa then do; /* categorizing uniform random variables into
outcome 'slots' using pa, pb, and pc */
vec[i,]=(1||1); /* and creating a matrix of said simulated outcomes */
end;
if u>pa then do;
```

```
if u<=pb then do;
vec[i,]=(1||0);
end;
if u>pb then do;
if u<=pc then do;
vec[i,]=(0||1);
end;
if u>pc then do;
vec[i,]=(0||0);
end;
end;
end;
/* End DO loop for simulations */
end;
```

Three-variable code for each technique above would be an extension of the code shown here.

A.3 Additional code for the two-group pre-/post-treatment simulation study for Section 2.5.1

```
/* Each subject has two or three repeated measures, depending on
whether two or three time points were used in the simulated study */
/* Subject identifier = uid, group identifier = group, time point
identifier (pre-treatment, post-treatment 1, or post-treatment 2
(if three-test case)) = prepost, also make sure the data are sorted
properly by prepost */
proc genmod data=data descending;
class outcome group(ref='1') prepost(ref='Pre') uid;
model outcome = group prepost group*prepost / dist=bin; /* logit link */
```

```
repeated subject=uid / corr=un corrw; /* corrw produces the working
correlation */
run;
```

Appendix B

SAS CODE RELEVANT TO CHAPTER 3

```
proc iml;
/* Create a function for the Plackett copula for finding the joint
probability */
start Plackett(u1,u2,psi);
e=psi-1;
pij=1/(2*e)*(1+e*(u1+u2)-((1+e*(u1+u2))**2-4*psi*e*u1*u2)**(1/2));
return(pij);
finish Plackett;
/* Set n, p1, p2, p3, OR12t, and OR23t before continuing */
/* Frechet bounds on the Odds Ratio - Chaganty and Joe (2006)
assuming the unstructured case */
/* Solve for the joint probabilities using the odds ratio and the
marginal probabilities in the Plackett copula */
p12=Plackett(p1,p2,OR12t);
p23=Plackett(p2,p3,OR23t);
/* Find limits on p13 */
p13L=max(0, p12+p23-p2, p1+p2+p3-p12-p23-1, p1+p3-1);
p13U=min(p1, p3, p1+p23-p12, p3+p12-p23);
/* Find limits on OR13 - Equation 9 from Chaganty and Joe (2006) */
OR13L=(p13L*(1-p1-p3+p13L))/((p1-p13L)*(p3-p13L));
if p13U=p1 | p13U=p3 then do; OR13U=100; inf=1; end;
else do; OR13U=(p13U*(1-p1-p3+p13U))/((p1-p13U)*(p3-p13U)); inf=0;
end;
```

```
if OR13L=0 then logLFB=-3;
else logLFB=log10(OR13L);
do logR=logLFB to log10(OR13U) by (log10(OR13U)-logLFB)/99;
R=10**(logR);
if 0.9999<=R & R <=1.0001 then Ra=1.01;
else Ra=R;
S13=((1+(p1+p3)*(Ra-1))**2+4*Ra*(1-Ra)*p1*p3)**(1/2);
p13=(1+(p1+p3)*(Ra-1)-S13)/(2*(Ra-1));
if (p13L <= p13 & p13 <= p13U) then do;
/* Find maximum and minimum probability for three successes, take
average for calculations */
p123l=max(0,p12+p13-p1,p12+p23-p2,p13+p23-p3);
p123u=min(p12,p13,p23,1-p1-p2-p3+p12+p13+p23);
p123=(p123l+p123u)/2;
p111=p123;
p110=p12-p123;
p101=p13-p123;
p011=p23-p123;
p100=p1-p12-p13+p123;
p010=p2-p12-p23+p123;
p001=p3-p13-p23+p123;
p000=1-p1-p2-p3+p12+p13+p23-p123;
/* Create the CDF of the distribution for categorizing a U(0,1)
random variable */
pa=p111;
pb=p111+p110;
pc=p111+p110+p101;
```

```
pd=p111+p110+p101+p011;
pe=p111+p110+p101+p011+p100;
pf=p111+p110+p101+p011+p100+p010;
pg=p111+p110+p101+p011+p100+p010+p001;
ph=p111+p110+p101+p011+p100+p010+p001+p000;
m=10000;
/* creating column vectors of zeroes with m rows */
cat=j(m,1,0);  /* Column for categorizing whether the estimated
correlation falls within the Frechet bounds */
adj12=j(m,1,0);
adj13=j(m,1,0);
adj23=j(m,1,0);
OR12=j(m,1,0); /* Columns for the estimated odds ratios */
OR13=j(m,1,0);
OR23=j(m,1,0);
ckv=j(m,1,0); /* Column for marking whether an observed variance is
zero */
do j=1 to m;
vec=j(n,3,0); /* creating nx3 matrix of zeroes */
uvec=j(n,1,0);
n1200=j(n,1,0); n1201=j(n,1,0); n1210=j(n,1,0); n1211=j(n,1,0);
n1300=j(n,1,0); n1301=j(n,1,0); n1310=j(n,1,0); n1311=j(n,1,0);
n2300=j(n,1,0); n2301=j(n,1,0); n2310=j(n,1,0); n2311=j(n,1,0);
do i=1 to n;
seed=47; /* random seed for calculating Uniform(0,1) random variable
*/
c=j(1,1,seed); /* 1x1 matrix seed */
```

```
/* Assigning multivariate binary outcomes to uniform random variable */
u=uniform(c); uvec[i,]=u; /* categorizing uniform random variables
into outcome 'slots' and creating a matrix of said simulated
outcomes*/  if u<=pa then vec[i,]=(1||1||1);
else if u<=pb then vec[i,]=(1||1||0);
else if u<=pc then vec[i,]=(1||0||1);
else if u<=pd then vec[i,]=(0||1||1);
else if u<=pe then vec[i,]=(1||0||0);
else if u<=pf then vec[i,]=(0||1||0);
else if u<=pg then vec[i,]=(0||0||1);
else if u<=ph then vec[i,]=(0||0||0);
/* Set up ennumeration for odds ratio calculation */
if vec[i,]=(0||0||0) | vec[i,]=(0||0||1) then n1200[i]=1; else
n1200[i]=0;
if vec[i,]=(0||1||0) | vec[i,]=(0||1||1) then n1201[i]=1; else
n1201[i]=0;
if vec[i,]=(1||0||0) | vec[i,]=(1||0||1) then n1210[i]=1; else
n1210[i]=0;
if vec[i,]=(1||1||0) | vec[i,]=(1||1||1) then n1211[i]=1; else
n1211[i]=0;
if vec[i,]=(0||0||0) | vec[i,]=(0||1||0) then n1300[i]=1; else
n1300[i]=0;
if vec[i,]=(0||0||1) | vec[i,]=(0||1||1) then n1301[i]=1; else
n1301[i]=0;
if vec[i,]=(1||0||0) | vec[i,]=(1||1||0) then n1310[i]=1; else
n1310[i]=0;
if vec[i,]=(1||0||1) | vec[i,]=(1||1||1) then n1311[i]=1; else
```

```
n1311[i]=0;
if vec[i,]=(0||0||0) | vec[i,]=(1||0||0) then n2300[i]=1; else
n2300[i]=0;
if vec[i,]=(0||0||1) | vec[i,]=(1||0||1) then n2301[i]=1; else
n2301[i]=0;
if vec[i,]=(0||1||0) | vec[i,]=(1||1||0) then n2310[i]=1; else
n2310[i]=0;
if vec[i,]=(0||1||1) | vec[i,]=(1||1||1) then n2311[i]=1; else
n2311[i]=0;
end;
/* Calculate Odds Ratios */
/* Adjustment:  +0.5 to all cells if any cell is zero, plus a marker
for the adjustment */
if n1200[+] = 0 | n1201[+] = 0 | n1210[+] = 0 | n1211[+] = 0 then do;
n1200a=n1200[+]+0.5;  n1201a=n1201[+]+0.5;  n1210a=n1210[+]+0.5;
n1211a=n1211[+]+0.5; adj12a=1; end;
else  do;  n1200a=n1200[+];  n1201a=n1201[+];  n1210a=n1210[+];
n1211a=n1211[+]; adj12a=0; end;
if n1300[+] = 0 | n1301[+] = 0 | n1310[+] = 0 | n1311[+] = 0 then do;
n1300a=n1300[+]+0.5;  n1301a=n1301[+]+0.5;  n1310a=n1310[+]+0.5;
n1311a=n1311[+]+0.5; adj13a=1; end;
else  do;  n1300a=n1300[+];  n1301a=n1301[+];  n1310a=n1310[+];
n1311a=n1311[+]; adj13a=0; end;
if n2300[+] = 0 | n2301[+] = 0 | n2310[+] = 0 | n2311[+] = 0 then do;
n2300a=n2300[+]+0.5;  n2301a=n2301[+]+0.5;  n2310a=n2310[+]+0.5;
n2311a=n2311[+]+0.5; adj23a=1; end;
else  do;  n2300a=n2300[+];  n2301a=n2301[+];  n2310a=n2310[+];
```

```
n2311a=n2311[+]; adj23a=0; end;
/* Denominators for odds ratios */  den12=(n1201a*n1210a);
den13=(n1301a*n1310a);
den23=(n2301a*n2310a);
/* Final odds ratio calculation */  OR12a=n1200a*n1211a/den12;
OR13a=n1300a*n1311a/den13;
OR23a=n2300a*n2311a/den23;
quit;
```

The common odds ratio case would be coded similarly, with different limits for p_{13}.

Appendix C

SAS CODE RELEVANT TO CHAPTER 4

C.1 Binomial Fréchet Bounds

```
proc iml;
/* Create a function to calculate the binomial cumulative probabilities
to be summed */
start pg(j,n,p);
sum1=0;
do i=j to n by 1;
addon=comb(n,i) * p**i * (1-p)**(n-i); /* Binomial pdf */
sum1=sum1+addon; /* Binomial cdf */
end;
return (sum1);
finish pg;
/* Set n1, p1, q1, n2, p2, and q2 before continuing */
/* Calculate EL and EU */
do i=1 to n1;
do j=1 to n2;
pr1=pg(i,n1,p1);
pr2=pg(j,n2,p2);
prob1=pr1+pr2-1;
add1=max(prob1,0);
EL=EL+add1;
add2=min(pr1,pr2);
```

```
EU=EU+add2;
end;
end;
/* Calculate Frechet bounds */
rhoL=(EL-n1*n2*p1*p2)/sqrt(n1*n2*p1*q1*p2*q2);
rhoU=(EU-n1*n2*p1*p2)/sqrt(n1*n2*p1*q1*p2*q2);
quit;
```

C.2 Negative Binomial Fréchet Bounds

```
proc iml;
/* Create a function to calculate the negative binomial cumulative
probabilities to be summed */
start pg(s,f,p);
sum1=0;
m=s-1;
/* Since the negative binomial has infinite support, first find the
left-tail cdf */
do k=0 to m;
a=k+f-1;
addon=comb(a,k) * p**f * (1-p)**k; /* Negative binomial pdf */
sum1=sum1+addon; /* Negative binomial cdf - left-tail */
end;
/* Subtract from 1 to find the right-tail cdf */
prob=1-sum1; /* Negative binomial cdf - right-tail */
return (prob);
finish pg;
/* Set r1, p1, q1, r2, p2, and q2 before continuing */
```

```
/* Find initial values for algorithm */
pg1v=pg(1,r1,p1);  pg2v=pg(1,r2,p2);  pg12maxm=max(pg1v+pg2v-1,0);
pg12minm=min(pg1v,pg2v);
/* Set initial values */
n=2; EL=pg12maxm; EU=pg12minm; rhoL=-1.1; rhoU=1.1;
/* Begin DO loop for finding EL, rhoL, EU, and rhoU */
/* Set criteria for stopping the algorithm */
do until (inc<0.000001 & rhoL>=-1 & rhoL<=rhoU & rhoU>=rhoL & rhoU<=1);
rhoL1=rhoL;
rhoU1=rhoU;
pg1=pg(n,r1,p1);
pg2=pg(n,r2,p2);
pg1v=pg1v//pg1;
pg2v=pg2v//pg2;
m=n-1;
/* Set up vectors for finding all combinations of i and n for the lower
bound.  Think of the possible combinations of i and j as a matrix.
Here, one row vector and column vector are added to the matrix per
iteration (making sure not to count the ''corner'' twice).  */
pg12maxr=j(1,m,0);
pg12maxc=j(n,1,0);
/* Set up vectors for finding all combinations of i and n for the
upper bound in the same manner as the lower bound */
pg12minr=j(1,m,0);
pg12minc=j(n,1,0);
/* Calculate maxima and minima to create each new element of the
current vector */
```

```
do i=1 to m;
pg12maxr[1,i]=max(pg1v[n]+pg2v[i]-1,0);
pg12minr[1,i]=min(pg1v[n],pg2v[i]);
end;
/* Calculate maxima and minima to create each new element of the
current vector */
do i=1 to n;
pg12maxc[i,1]=max(pg1v[i]+pg2v[n]-1,0);
pg12minc[i,1]=min(pg1v[i],pg2v[n]);
end;
/* Sum the elements of the maximum vectors and the minimum vectors
separately */
pg12maxm=sum(pg12maxr,pg12maxc);
pg12minm=sum(pg12minr,pg12minc);
/* Add sums to current estimates of EL and EU */
EL=EL+pg12maxm;
EU=EU+pg12minm;
/* Calculate rhoL and rhoU based on estimates of EL and EU */
rhoL=(EL-r1*r2*(q1*q2)/(p1*p2))/sqrt(r1*r2*(q1*q2)/((p1*p2)**2));
rhoU=(EU-r1*r2*(q1*q2)/(p1*p2))/sqrt(r1*r2*(q1*q2)/((p1*p2)**2));
/* Calculate the maximum of the differences between the previous
rhoL and rhoU to compare to the specified tolerance */
inc1=abs(rhoL-rhoL1);
inc2=abs(rhoU-rhoU1);
inc=max(inc1,inc2);
n=n+1;
end;
```

```
quit;
```

C.3 NHANES Sample Analysis

```
/* Each subject has two repeated measures, one for environmental
phenols, and one for environmental pesticides.  */
/* Subject identifier = uid, Gender = gender, type = type of analytes
(phenols or pesticides) result = number of results >75th percentile,
trials = total number of phenols (8) or pesticides (5) measured */
proc genmod data=analysis;
class gender uid type;
model result/trials = gender type gender*type / dist=bin type3; /*
results/trials indicates binomial data, logit link used, Type 3
Chi-square tests requested */
repeated subject=uid / corr=un corrw; /* corrw produces the working
correlation */
run;
```

REFERENCES

[1] N. Rao Chaganty and Harry Joe. "Range of correlation matrices for dependent Bernoulli random variables". In: *Biometrika* 93.1 (2006), pp. 197–206.

[2] N. Rao Chaganty and Deepak Mav. "Estimation methods for analyzing longitudinal data occurring in biomedical research". In: *Computational Methods in Biomedical Research*. Ed. by R.Khattree and D.Naik. Chapman and Hall/CRC Press.

[3] Centers for Disease Control, Prevention (CDC), and National Center for Health Statistics (NCHS). *National Health and Nutrition Examination Survey Data - Demographics*. 2009-2010. URL: `http://wwwn.cdc.gov/nchs/nhanes/2009-2010/DEMO_F.XPT`.

[4] Centers for Disease Control, Prevention (CDC), and National Center for Health Statistics (NCHS). *National Health and Nutrition Examination Survey Data - Environmental Pesticides*. 2009-2010. URL: `http://wwwn.cdc.gov/nchs/nhanes/2009-2010/PP_F.XPT`.

[5] Centers for Disease Control, Prevention (CDC), and National Center for Health Statistics (NCHS). *National Health and Nutrition Examination Survey Data - Environmental Phenols*. 2009-2010. URL: `http://wwwn.cdc.gov/nchs/nhanes/2009-2010/EPH_F.XPT`.

[6] Lawrence J. Emrich and Marion R. Piedmonte. "A Method for Generating High-Dimensional Multivariate Binary Variates". In: *The American Statistician* 45.4 (1991), pp. 302–304.

[7] Maurice R. Fréchet. "Généralisations du théorème des probabilités totales". In: *Fundamenta Mathematica* 25 (1935), pp. 379–387.

[8] Maurice R. Fréchet. "Sur les tableaux de corrélation dont les marges sont données". In: *Annales de l'Université de Lyon. Section A: Sciences mathématiques et astronomie* 45 (1951), pp. 53–77.

[9] Kung-Yee Liang and Scott L. Zeger. “Longitudinal data analysis using generalized linear models”. In: *Biometrika* 73.1 (1986), pp. 13–22.

[10] Roy T. Sabo and N. Rao Chaganty. “What can go wrong when ignoring correlation bounds in the use of generalized estimating equations”. In: *Statistics in Medicine* 29 (2010), pp. 2501–2507.

[11] Scott L. Zeger and Kung-Yee Liang. “Longitudinal data analysis for discrete and continuous outcomes”. In: *Biometrics* 42.1 (1986), pp. 121–130.

VITA

Mary Emilia Haynes was born on May 23, 1987, in Macomb County, Michigan, and is an American citizen. She graduated from Macomb Christian Schools, Warren, Michigan in 2005. She received her Bachelor of Arts in Mathematics Foundations from Bryan College, Dayton, Tennessee in 2009. During her time at Virginia Commonwealth University in pursuit of her PhD, she interned with the biostatistics department of PharPoint Research, Inc. in Durham, North Carolina.